Actes de la 14e conférence sur
le Traitement Automatique des Langues Naturelles
(communications affichées et démonstrations)

Actes de la 11e
Rencontre des Étudiants Chercheurs en Informatique
pour le Traitement Automatique des Langues
(communications affichées)

Actes de l'atelier TALN
Formalismes syntaxiques de haut niveau

Actes de l'atelier TALN
Reconstruire la langue dans les communications
alternatives et augmentées

Conception graphique de l'affiche de la conférence : Benoît COLAS (Université Toulouse-le-Mirail, CPRS-UMS 838). Couverture (d'après l'affiche de la conférence) : Ludovic CHACUN (Institut de Recherche en Informatique de Toulouse, UMR 5505). Composition et mise en page : Dominique LONGIN (Institut de Recherche en Informatique de Toulouse, UMR 5505). Impression : Société Générale d'Impression (sgi31@wanadoo.fr).

 ISBN : 2-9520326-9-6

Table des matières

TALN-2007

5 au 8 juin 2007, Toulouse, France

Actes de la 14e conférence sur
le TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES
(communications affichées)

Éditeurs scientifiques
Nabil HATHOUT et Philippe MULLER

Organisation de la conférence
CLLE-ERSS (UMR 5263) & IRIT (UMR 5505)

Sous l'égide de l'ATALA
(Association pour le Traitement Automatique des Langues)

Comité d'organisation

Nathalie AUSSENAC-GILLES	*(CNRS, IRIT)*
Farah BENAMARA	*(Université Paul Sabatier, IRIT)*
Jean-Léon BOURAOUI	*(Université Paul Sabatier, IRIT)*
Didier BOURIGAULT	*(CNRS & Université Toulouse-Le-Mirail, CLLE)*
Véronique DEBATS	*(CNRS, IRIT)*
Fabrice ÉVRARD	*(Institut National Polytechnique, IRIT)*
Cécile FABRE	*(Université Toulouse-Le-Mirail, CLLE)*
Edith GALY	*(Université Toulouse-Le-Mirail, CLLE)*
Bruno GAUME	*(Université Toulouse-Le-Mirail, CLLE)*
Nabil HATHOUT*	*(CNRS, Université Toulouse-Le-Mirail, CLLE)*
Dominique LONGIN	*(CNRS, IRIT)*
Josiane MOTHE	*(Université Paul Sabatier, IRIT)*
Philippe MULLER*	*(Université Paul Sabatier, IRIT)*
Sylwia OZDOWSKA	*(Université Toulouse-Le-Mirail, CLLE)*
Patrick SAINT-DIZIER	*(CNRS, IRIT)*
Frank SAJOUS	*(Université Toulouse-Le-Mirail, CLLE)*
Ludovic TANGUY	*(Université Toulouse-Le-Mirail, CLLE)*
Laure VIEU	*(CNRS, IRIT)*

Comité de programme

Salah AIT-MOKHTAR	*(Xerox Research Centre Europe, XRCE)*
Nathalie AUSSENAC-GILLES	*(CNRS, IRIT)*
Philippe BLACHE	*(CNRS, LPL)*
Yves BESTGEN	*(Université catholique de Louvain, FNRS)*
Didier BOURIGAULT	*(CNRS & Université Toulouse-Le-Mirail, CLLE)*
Jean CAELEN	*(Université Joseph Fourier, CLIPS-IMAG)*
Vincent CLAVEAU	*(Université de Rennes 1, IRISA)*
Beatrice DAILLE	*(Université de Nantes, LINA)*
Laurence DANLOS	*(Université Paris 7, Lattice)*
Éric DE LA CLERGERIE	*(INRIA, Atoll)*
Cédrick FAIRON	*(Université Catholique de Louvain)*
Claire GARDENT	*(CNRS, LORIA)*
Nabil HATHOUT*	*(CNRS & Université Toulouse-Le-Mirail, CLLE)*
Sylvain KAHANE	*(Université Paris 10, Modyco)*
Philippe LANGLAIS	*(Université de Montréal, RALI)*
Dominique LAURENT	*(Synapse Développement)*
Piet MERTENS	*(Katholieke Universiteit Leuven, Faculteit Letteren)*
Detmar MEURERS	*(Ohio State University, CLLT)*
Philippe MULLER*	*(Université Paul Sabatier, IRIT)*
Fiammetta NAMER	*(Université de Nancy 2, ATILF)*
Anne NICOLLE	*(Université de Caen, GREYC)*
Patrick PAROUBEK	*(CNRS, LIMSI)*
Jean-Marie PIERREL	*(Nancy Université & CNRS, ATILF)*
Owen RAMBOW	*(Université de Columbia, CCLS)*
Sophie ROSSET	*(CNRS, LIMSI)*
François YVON	*(ENST, GET)*
Pierre ZWEIGENBAUM	*(CNRS, LIMSI ; CRIM-INALCO)*

* Président

Comité scientifique

Ramzi ABBES
Salah AIT-MOKHTAR
Pascal AMSILI
Carlos ARECES
Denis BECHET
Patrice BELLOT
Yves BESTGEN
Hervé BLANCHON
Pierrette BOUILLON
Didier BOURIGAULT
Jean CAELEN
Lionel CLÉMENT
Laurence DANLOS
Éric DE LA CLERGERIE
Hervé DÉJEAN
Marc EL-BÈZE
Patrice ENJALBERT
Cédrick FAIRON
Thierry FONTENELLE
Claire GARDENT
Brigitte GRAU
Marie-Laure GUÉNOT
Lapalme GUY
Nabil HATHOUT
Diana INKPEN
Sylvain KAHANE
Daniel KAYSER
Mathieu LAFOURCADE
Éric LAPORTE
Alain LECOMTE
Denis MAUREL
Detmar MEURERS
Laura MONCEAUX
Michel MOREL
Philippe MULLER
Alexis NASR
Patrick PAROUBEK
Sébastien PAUMIER
Marie-Paule PERY-WOODLEY
Sylvain POGODALLA
Laurent PREVOT
Owen RAMBOW
Laurent ROMARY
Jean ROYAUTÉ
Gérard SABAH
Pascale SÉBILLOT
Jesse TSENG
Mathieu VALETTE
François YVON
Pierre ZWEIGENBAUM

Anne ABEILLE
Susanne ALT
Jean-Yves ANTOINE
Nathalie AUSSENAC-GILLES
Núria BEL
Romaric BESANÇON
Philippe BLACHE
Malek BOUALEM
Philippe BOULA DE MAREÜIL
Ilana BROMBERG
Vincent CLAVEAU
Beatrice DAILLE
Gaël DE CHALENDAR
Claude DE LOUPY
Marc DYMETMAN
Chantal ENGUEHARD
Jacquey EVELYNE
Olivier FERRET
Nuria GALA
Natalia GRABAR
Gregory GREFENSTETTE
Bruno GUILLAUME
Thierry HAMON
Nicolas HERNANDEZ
Christine JACQUIN
Mouna KAMEL
Olivier KRAIF
Philippe LANGLAIS
Dominique LAURENT
Yves LEPAGE
Piet MERTENS
Jean-Luc MINEL
Richard MOOT
Emmanuel MORIN
Fiammetta NAMER
Anne NICOLLE
Patrick SAINT-DIZIER
Guy PERRIER
Jean-Marie PIERREL
Thierry POIBEAU
Violaine PRINCE
Paul RAYSON
Sophie ROSSET
C. Anton RYTTING
Benoît SAGOT
François TROUILLEUX
Agnès TUTIN
Tristan VANRULLEN
Michael ZOCK

Session
Communications affichées

Désambiguïsation lexicale automatique : sélection automatique d'indices

Laurent AUDIBERT
Laboratoire d'Informatique de l'université Paris-Nord (LIPN)
99, avenue Jean-Baptiste Clément – 93430 Villetaneuse, France
laurent.audibert@lipn.univ-paris13.fr

Résumé. Nous exposons dans cet article une expérience de sélection automatique des indices du contexte pour la désambiguïsation lexicale automatique. Notre point de vue est qu'il est plus judicieux de privilégier la pertinence des indices du contexte plutôt que la sophistication des algorithmes de désambiguïsation utilisés. La sélection automatique des indices par le biais d'un algorithme génétique améliore significativement les résultats obtenus dans nos expériences précédentes tout en confortant des observations que nous avions faites sur la nature et la répartition des indices les plus pertinents.

Abstract. This article describes an experiment on automatic features selection for word sense disambiguation. Our point of view is that word sense disambiguation success is more dependent on the features used to represent the context in which an ambiguous word occurs than on the sophistication of the learning techniques used. Automatic features selection using a genetic algorithm improves significantly our last experiment bests results and is consistent with the observations we have made on the nature and space distribution of the most reliable features.

Mots-clés : désambiguïsation lexicale automatique, corpus sémantiquement étiqueté, cooccurrences, sélection d'indices, algorithmes génétiques.

Keywords: word sense disambiguation, sense tagged corpora, cooccurrences, features selection, genetic algorithms.

1 Introduction

La plupart des mots ont plusieurs significations. La *désambiguïsation lexicale* consiste à choisir la bonne signification d'un mot polysémique dans un contexte donné. Cette opération est utile ou indispensable pour la plupart des applications de traitement automatique des langues : recherche d'information, traduction automatique, reconnaissance de la parole, etc. (Ide & Véronis, 1998). La campagne d'évaluation trisannuel SensEval (Edmonds, 2002) atteste de l'importance de cette tâche.

La désambiguïsation lexicale s'effectue toujours en utilisant l'information présente dans le contexte du mot à désambiguïser. Cette information peut être enrichie par un certain nombre d'annotations (étiquette morphosyntaxique, lemmatisation, etc.). Il n'est cependant pas pos-

sible d'utiliser toute l'information disponible car elle est bien trop importante et bruitée. Il faut donc se focaliser sur un certain nombre d'indices. Le choix de ces indices, déterminé par ce que nous appelons des critères de désambiguïsation lexicale, est primordial et constitue un enjeu important dans le domaine de la désambiguïsation lexicale automatique (Bruce *et al.*, 1996; Ng & Zelle, 1997; Pedersen, 2001b).

Notre approche s'inscrit dans celles qui utilisent des techniques de classification supervisée sur un corpus lexicalement désambiguïsé. Dans ce type d'approche, de nombreux travaux cherchent à améliorer la précision de la désambiguïsation en améliorant les techniques de classification. Le choix des indices utilisés est généralement déterminé plus ou moins arbitrairement par la connaissance, l'expérience et l'intuition du chercheur. Peu de travaux avaient étudié systématiquement l'impact du choix des indices utilisés sur la précision de la désambiguïsation. Pour cette raison, nous avons présenté une étude des critères de désambiguïsation sémantique automatique (Audibert, 2003a) basés sur les unigrammes (*i.e.* cooccurrences de mots isolés). Nous avons complété cette étude en explorant des indices basés sur des bigrammes et des trigrammes (Audibert, 2004). Dans ces travaux, les critères étudiés étaient *homogènes* dans le sens où ils étaient constitués d'indices de même nature : par exemple, soit des lemmes, soit des étiquettes morphosyntaxiques, mais pas une combinaison des deux.

Dans le présent article, nous présentons, dans un premier temps, une petite étude comparative de différents algorithmes de classification. Nous nous intéressons ensuite à la sélection automatique des meilleurs indices du contexte pour former des critères de désambiguïsation hétérogènes sur lesquels un algorithme de classification peut s'appuyer efficacement pour effectuer de la désambiguïsation lexicale. Ce travail s'appuie toujours sur les 60 mots cibles (20 noms, 20 adjectifs et 20 verbes) des travaux précédents (Audibert, 2003a; Audibert, 2004).

2 Corpus, indices et critères

2.1 Corpus

Notre corpus de travail est composé de textes de genres variés et comporte 6 468 522 mots. Il a été constitué dans le cadre du projet *SyntSem* qui vise à produire un corpus français d'amorçage étiqueté au niveau morphosyntaxique, lemmatisé et comportant un étiquetage syntaxique peu profond ainsi qu'un étiquetage lexical de 60 mots-cibles sélectionnés pour leur caractère fortement polysémique (Véronis, 1998). Ces 60 mots-cibles, qui totalisent 53796 occurrences dans le corpus, sont également répartis en 20 noms, 20 adjectifs et 20 verbes et sont détaillés dans le tableau 1.

L'une des difficultés majeures de l'étiquetage sémantique automatique réside dans l'inadéquation des dictionnaires traditionnels (Véronis, 2001) ou dédiés (Palmer, 1998) pour cette tâche. Pour remédier à ce problème, l'équipe DELIC [1] a entrepris la construction d'un dictionnaire distributionnel en se basant sur un ensemble de critères différentiels stricts (Reymond, 2001). C'est ce dictionnaire qui a été utilisé pour étiqueter les occurrences des 60 mots-cibles du projet *SyntSem*. Dans ce dictionnaire, le nombre de lexies par vocable est important car il inclut les locutions figées ou composées comme *mettre sur pied*, *mettre à pied*, *pied de nez*, etc.

Un consensus semble émerger selon lequel l'étiquetage morphosyntaxique, et plus particuliè-

[1] Équipe DELIC, Université de Provence, 29 Avenue Robert SCHUMAN, 13621 Aix-en-Provence Cedex 1.

Noms				Adjectifs				Verbes			
Vocable	**freq**	**lex**	**H**	**Vocable**	**freq**	**lex**	**H**	**Vocable**	**freq**	**lex**	**H**
barrage	92	5	1,18	correct	116	5	1,81	couvrir	518	21	3,25
restauration	104	5	1,85	sain	129	10	2,45	importer	576	8	2,57
suspension	110	5	1,50	courant	168	4	0,63	parvenir	653	8	2,31
détention	112	2	0,85	régulier	181	11	2,54	exercer	698	8	1,52
lancement	138	5	0,99	frais	182	18	3,10	conclure	727	16	2,36
concentration	246	6	1,98	secondaire	195	5	1,69	arrêter	913	15	2,97
station	266	8	2,58	strict	220	9	2,23	ouvrir	919	41	3,80
vol	278	10	2,20	exceptionnel	226	3	1,45	poursuivre	978	16	2,71
organe	366	6	2,24	utile	359	9	2,39	tirer	1001	47	3,88
compagnie	412	12	1,62	vaste	368	6	2,08	conduire	1082	15	2,28
constitution	422	6	1,64	sensible	425	11	2,63	entrer	1210	38	3,65
degré	507	18	2,47	traditionnel	447	2	0,49	connaître	1635	16	2,24
observation	572	3	0,68	populaire	457	5	2,02	rendre	1985	27	2,88
passage	601	19	2,70	biologique	475	4	0,55	comprendre	2136	13	2,76
solution	880	4	0,44	clair	556	20	3,10	présenter	2140	18	2,56
économie	930	10	2,16	historique	620	3	0,67	porter	2328	59	4,01
pied	960	62	3,55	sûr	645	14	2,61	répondre	2529	9	0,99
chef	1133	11	1,47	plein	844	35	3,99	passer	2547	83	4,49
formation	1528	9	1,66	haut	1016	29	3,46	venir	3788	33	3,21
communication	1703	13	2,44	simple	1051	14	2,14	mettre	5095	140	3,65
Moyenne	568	14,2	1,9	**Moyenne**	434,4	14,1	2,3	**Moyenne**	1687,6	47,4	3,1

TAB. 1 – Fréquence moyenne des occurrences des vocables (**freq**), nombre moyen de lexies (**lex**) et entropie de la répartition des occurrences sur les lexies (**H**).

rement la levée de l'ambiguïté sur la catégorie grammaticale des vocables, n'est pas du ressort de la désambiguïsation lexicale (Kilgarriff, 1997; Ng & Zelle, 1997). Nous avons confié l'étiquetage morphosyntaxique de notre corpus au logiciel *Cordial Analyseur* (développé par la société Synapse Développement), qui offre une lemmatisation et un étiquetage morphosyntaxique d'une exactitude satisfaisante (Valli & Véronis, 1999).

jeton	**lemme**	**ems**	**smallems**	**lexie**
pouvait	pouvoir	VINDI3S	VCON	
mettre	mettre	VINF	VINF	1.12.7
fin	fin	NCFS	NCOM	
à	à	PREP	PREP	
la	le	DETDFS	DET	
pratique	pratique	NCFS	NCOM	
des	de	DETDPIG	DET	
détentions	détention	NCFP	NCOM	1

TAB. 2 – Extrait du corpus *SyntSem*

Le Tableau 2 présente un extrait du corpus *SyntSem*. Il permet de visualiser l'ensemble des étiquettes que possède un mot. C'est l'information de ces étiquettes que nous utilisons dans nos critères de désambiguïsation lexicale.

2.2 Indices et critères

Nous désignons par le terme d'*indice* une source potentielle d'information pouvant participer à la levée de l'ambiguïté d'un mot cible dont nous cherchons la bonne lexie. Un indice peut être le lemme du mot qui précède par exemple. Un *critère* est simplement la donnée d'un ensemble d'indices.

Nous avons étudié une grande variété de critères dans (Audibert, 2004). Les noms de ces critères précisent leur nature et sont de la forme `[P1|P2|P3|P4]`. Le paramètre `P1` indique si le critère considère des unigrammes (`P1=1gr`), des bigrammes (`P1=2gr`) ou des trigrammes (`P1=3gr`) ; un n-gramme étant la juxtaposition de n mots. Le paramètre `P2` indique si l'on regarde la forme brute des mots (`P2=jeton`), leur lemme (`P2=lemme`), leur étiquette morphosyntaxique (`P2=ems`) ou leur étiquette morphosyntaxique simplifiée (`P2=smallems`). Le paramètre `P3` indique si les mots considérés sont différenciés par leur position (`P3=ordonne`), différenciés suivant qu'ils appartiennent au contexte droit ou gauche (`P3=differencie`), ou non différenciés (`P3=non-ordonne`). Enfin, le paramètre `P4` indique si le critère considère tous les mots (`P4=mot`) ou seulement les mots pleins (`P4=mot-plein`). Nous qualifions ces critères de *critères homogènes* dans la mesure où l'ensemble des indices de désambiguïsation sont de la même nature puisque entièrement déterminés par l'instanciation des quatre paramètres.

3 Comparaison de différents algorithmes de désambiguïsation

Dans cette expérience, nous comparons différents algorithmes de classification supervisée en utilisant un critère assez standard constitué du lemme des mots en tenant compte de leur position (*i.e.* `[1gr|lemme|ordonne|mot]`) dans une fenêtre de ± 3 mots. Les algorithmes de classification évalués sont les suivants :

MAJ est un classifieur qui retourne toujours la lexie la plus fréquente ; nous l'utilisons comme borne inférieure à la précision de la désambiguïsation ;

PCM est un algorithme basé sur une liste de décisions, proche de celui utilisé par (Yarowsky, 1994) et détaillé dans (Audibert, 2003a) ;

NB est notre implémentation du classifieur Naïf de Bayes ;

KPPV est une implémentation élémentaire d'un classifieur du type k plus proches voisins ;

PEBLS est classifieur du type k plus proches voisins possédant une métrique bien plus sophistiqué que celle de `KPPV` ;

NBW est l'implémentation du projet Weka du classifieur Naïf de Bayes ;

C45W est l'implémentation du projet Weka du classifieur C45.

Le tableau 3 montre les résultats de cette expérience comparative. Dans toutes les expériences de désambiguïsation de cet article, toutes les occurrences reçoivent une classification. Le rappel étant égal à la précision dans ce cas, nous ne mentionnons que la précision obtenue.

Les temps d'exécution des deux algorithmes du projet Weka que nous avons utilisés (`NBW` et `C45W`) sont rédhibitoires pour nos expériences. Les raisons de ces temps d'exécution sont, ou peuvent être, la non optimisation de l'implémentation, l'utilisation du langage java et le format,

	MAJ	PCM	NB	KPPV	PEBLS	NBW	C45W
Précision	$42,9\%$	$72,3\%$	$74,5\%$	$65,5\%$	$70,9\%$	$58,2\%$	$74,6\%$
Intervalle de confiance		$\pm 0,38\%$	$\pm 0,37\%$	$\pm 0,40\%$	$\pm 0,38\%$	$\pm 0,42\%$	$\pm 0,37\%$
Temps	3s	3s	5s	26mn	2h33mn	1h47mn	35h43mn

TAB. 3 – Comparaison de la précision, avec intervalle de confiance de l'estimation à 95%, et des temps d'exécution de différents algorithmes de classification.

peu adapté au problème, de la représentation des données d'apprentissage. Le temps d'exécution du classifieur `PEBLS` est également bien trop important et est une conséquence de la complexité de la métrique utilisée.

Les classifieurs `NB` et `PCM`, nécessitent tous deux des estimations de probabilités. En raison des observations souvent rares et parfois nulles qui interviennent dans ces estimations, nous utilisons la m-estimation (Cussens, 1993) plutôt que l'estimation classique des probabilités. Cette différence explique certainement l'écart de performance des classifieurs `NB` et `NBW`.

Nous pouvons tirer deux enseignements de cette expérience. La premier est qu'il est souvent difficile et parfois préjudiciable d'utiliser un algorithme de classification comme une boîte noire (cf. la comparaison entre `NB` et `NBW`). Le second est que la complexité et la sophistication des algorithmes de classification n'apportent pas forcément un gain important pour notre tâche (cf. la comparaison entre `NB`, `PEBLS` et `C45`). Actuellement, des gains bien plus importants sont à attendre des indices fournis aux classifieurs plutôt que des classifieurs eux-mêmes.

4 Sélection automatique des indices

4.1 Méthodologie

En prenant tous les indices générés par tous les critères homogènes `[P1|P2|P3|P4]` correspondants aux différentes instanciations possibles des quatre paramètres `P1` à `P4`, et en considérant une fenêtre de ± 12 mots, nous obtenons 3 (`1gr`, `2gr` ou `3gr`) $\times 4$ (`jeton`, `lemme`, `ems` ou `smallems`) $\times 3$ (`ordonne`, `differentie` ou `non-ordonne`) $\times 2$ (`mot` ou `mot-plein`) $\times 24$ (contexte de ± 12 mots[2]) soit 1728 indices différents.

En réduisant la taille du contexte considéré, nous avons généré un deuxième jeu d'indices réduit à 888 indices. Dans ce jeu d'indices, la taille du contexte pour les critères basés sur les étiquettes `lemme` et `jeton` et composés d'unigrammes (respectivement de bigrammes et trigrammes) est de ± 6 mots (respectivement ± 8 et ± 10), et pour les critères basés sur les étiquettes `ems` et `smallems` et composés d'unigrammes (respectivement de bigrammes et trigrammes) est de ± 4 mots (respectivement ± 5 et ± 6).

La question est de savoir quels indices retenir, parmi les 1728 du premier jeu d'indices ou parmi les 888 du second, pour former un critère hétérogène efficace pour la levée de l'ambiguïté. Pour représenter un critère nous utilisons une chaîne de bits, appelée un génome, composée de 1728

[2] Le calcul est ici simplifié, en réalité, le nombre d'indices considérés sans sortir du contexte de ± 12 mots est de $12 + 1 + 12 = 25$ pour les unigrammes, $12 + 1 + 11 = 24$ pour les bigrammes et $12 + 1 + 10 = 23$ pour les trigrammes ce qui fait bien 24 en moyenne.

bits pour le premier jeu et de 888 bits pour le second. La valeur de chaque bit permet de préciser si l'indice associé est retenu ou pas. Un génome caractérise donc un critère (une sélection d'indices) hétérogène (tous les indices ne sont pas forcément de la même nature). En raison de la complexité combinatoire de notre problème d'optimisation de sélection d'indices, il n'existe pas de méthode exacte pour le résoudre en un temps raisonnable. Il faut donc se contenter de solutions approchées que nous obtenons en utilisant deux techniques classiques d'optimisation : les algorithmes gloutons et les algorithmes génétiques[3]. Le principe de l'algorithme glouton est de rechercher le meilleur indice pris individuellement, puis de chercher quel indice lui associer pour améliorer au maximum la précision, et ainsi de suite jusqu'à ne plus obtenir d'amélioration. Les algorithmes génétiques, quant à eux, tentent de mettre en œuvre le principe de la sélection naturelle (croisements et mutations) sur des populations de solutions potentielles (*i.e* des génomes) et se rapprochent de la solution au cours de générations successives.

Pour mettre en œuvre ces techniques, le corpus de départ est scindé en deux sous-corpus. Le premier sous-corpus contient 60% des exemples d'apprentissage. Il est utilisé dans un premier temps pour effectuer la sélection des indices en utilisant l'algorithme glouton ou l'algorithme génétique. Cette sélection se fait en générant une famille de génomes en suivant les règles propres à l'algorithme glouton ou génétique. L'évaluation de la performance de chacun des génomes (*i.e.* sous-ensemble d'indices) est réalisée par l'estimation de la précision obtenue par le classifieur `NB` en utilisant une méthode d'évaluation croisée k fois (avec $k = 10$) toujours sur ce même sous-corpus. Cette méthode est coûteuse en temps de calcul, mais permet l'évaluation des critères (*i.e.* des génomes) sur la totalité du sous-corpus. Une nouvelle génération de génomes est ensuite calculée en fonction de la génération précédente et des règles de l'algorithme glouton ou génétique. L'expérience est répétée tant que des génomes plus performants émergent des générations successives.

Les indices sélectionnés par le génome obtenant la meilleure performance constituent un critère hétérogène utilisé pour l'apprentissage du classifieur `NB` sur la totalité du sous-corpus contenant 60% des exemples. Le deuxième sous-corpus, qui contient 40% des exemples d'apprentissage, est enfin utilisé pour estimer la précision de désambiguïsation obtenue par le classifieur `NB` précédemment entraîné.

Cette expérience a été conduite d'un côté sur chacun des vocables indépendamment (*i.e.* un génome est sélectionné pour chacun des vocables) et d'un autre côté par catégorie grammaticale (*i.e.* un unique génome est sélectionné pour les 20 vocables d'une catégorie). L'expérience par catégorie grammaticale n'a pas été menée pour le jeu contenant 1728 indices en raisons des temps de calcul déjà de l'ordre de la dizaine de jours pour le jeu contenant 888 indices. Le tableau 4 rend compte des résultats de notre expérience.

4.2 Résultats des différentes expériences de sélection

La lecture du tableau 4 permet d'observer immédiatement que chacune des expériences de sélection automatique des indices à permis de surpasser la précision obtenue par le meilleur critère homogène identifié dans (Audibert, 2004).

Nous avons systématiquement obtenu de meilleurs résultats en sélectionnant les indices avec l'algorithme génétique plutôt qu'avec l'algorithme glouton qui est incapable de se sortir d'un

[3] Ce type d'approche n'est pas original, par exemple (Daelemans *et al.*, 2003) montrent comment obtenir une amélioration significative des performances en réalisant une optimisation simultanée des paramètres de l'algorithme d'apprentissage et de la sélection des indices en utilisant justement des algorithmes génétiques.

	Noms			Adjectifs			Verbes			Moyenne		
	P (%)	Am.	ICA	P (%)	Am.	ICA	P (%)	Am.	ICA	P (%)	Am.	ICA
Baseline (MAJ)	$57,2$			$46,3$			$37,2$			$42,9$		
Critère homogène	$81,4$	$0,0$		$75,1$	$0,0$		$72,3$	$0,0$		$74,7$	$0,0$	
Glou/Voc (1728)	$82,5$	$1,0$	$\pm 1,6$	$75,7$	$0,5$	$\pm 2,0$	$74,3$	$2,0$	$\pm 1,1$	$76,3$	$1,6$	$\pm 0,8$
Géné/Voc (1728)	$83,5$	$2,1$	$\pm 1,6$	$75,9$	$0,7$	$\pm 2,0$	$75,1$	$2,8$	$\pm 1,0$	$77,0$	$2,3$	$\pm 0,8$
Glou/Voc (888)	$82,9$	$1,5$	$\pm 1,6$	$75,7$	$0,6$	$\pm 2,0$	$75,0$	$2,7$	$\pm 1,0$	$76,8$	$2,1$	$\pm 0,8$
Géné/Voc (888)	$85,3$	$3,9$	$\pm 1,5$	$77,3$	$2,2$	$\pm 2,0$	$77,3$	$5,0$	$\pm 1,0$	$79,0$	$4,3$	$\pm 0,8$
Glou/Cat (888)	$83,7$	$2,3$	$\pm 1,6$	$75,9$	$0,7$	$\pm 2,0$	$76,8$	$4,5$	$\pm 1,0$	$78,1$	$3,4$	$\pm 0,8$
Géné/Cat (888)	$85,9$	$4,4$	$\pm 1,5$	$78,2$	$3,1$	$\pm 2,0$	$77,7$	$5,4$	$\pm 1,0$	$79,5$	$4,8$	$\pm 0,8$

TAB. 4 – Précision d'un critère hétérogène constitué par sélection automatique des indices. La ligne *Baseline (MAJ)* donne la précision obtenue par l'algorithme retournant systématiquement la lexie majoritaire. La ligne *Critère homogène* donne la précision obtenue par le meilleur critère homogène, c'est-à-dire le critère `[2gr|lemme|differencie|mot]`) avec une taille de fenêtre de ± 4 mots pour les noms et les verbes et ± 3 mots pour les adjectifs. Dans les lignes suivantes, *Glou* signifie que la technique de sélection d'indices utilisée est de type algorithme glouton tandis que *Géné* signifie que la technique utilisée est de type algorithme génétique. *Voc* signifie que la sélection d'indices est indépendante pour chacun des vocables et *Cat* qu'elle est commune aux 20 vocables de la catégorie grammaticale. *(1728)* et *(888)* précisent la taille du jeux d'indices de l'expérience. La colonne *P (%)* donne la précision obtenue en pourcentage, *Am.* l'amélioration réalisée par rapport à la précision du meilleur critère homogène (ligne *Critère homogène*) et *ICA* l'intervalle de confiance à 95% de l'amélioration réalisée (l'intervalle de confiance de la précision étant bien inférieur).

minimum local. Dans nos expériences, l'algorithme glouton sélectionne environ 20 indices. D'un autre côté, un algorithme génétique est capable, par définition, de se sortir d'un minimum local. Cependant, il ne garantit pas que tous les indices sélectionnés sont utiles et il sélectionne, dans nos expériences, environ 180 indices.

Un autre phénomène qui ressort de la lecture de ces résultats est que le jeu d'indices qui n'en contient que 888 permet d'aboutir à de meilleurs résultats que le jeu d'indices en contenant 1728. Les deux raisons de ce comportement sont la taille du corpus d'apprentissage, probablement trop faible pour une telle quantité d'indices, et le piège du surapprentissage sensible dans notre approche.

De manière surprenante, nous obtenons de meilleurs résultats en opérant la sélection sur l'ensemble d'une catégorie grammaticale plutôt que sur chacun des vocables pris individuellement. Opérer la sélection sur l'ensemble d'une catégorie grammaticale permet de limiter le phénomène de surapprentissage et d'augmenter le nombre d'exemples sur lesquels se fait la sélection. Le gain obtenu par la limitation du phénomène de surapprentissage et l'augmentation du nombre d'exemples est ici supérieur à celui obtenu par l'ajustement de la sélection des indices individuellement pour chaque vocable.

L'accord entre plusieurs annotateurs (*ITA* pour *InTer-annotator Agreement* en anglais) a été estimé à 96.4% (Audibert, 2003b) sur notre corpus. La précision moyenne de $79,5\%$ obtenue en effectuant une sélection automatique des indices permet de gagner $4,8pt$ (avec un intervalle de confiance de $\pm 0,8$) sur la précision obtenue par le meilleur critère homogène, ce qui correspond à 22% de l'écart avec la borne maximale estimée. Il s'agit donc d'une amélioration très substantielle.

4.3 Forme et répartition des indices sélectionnés

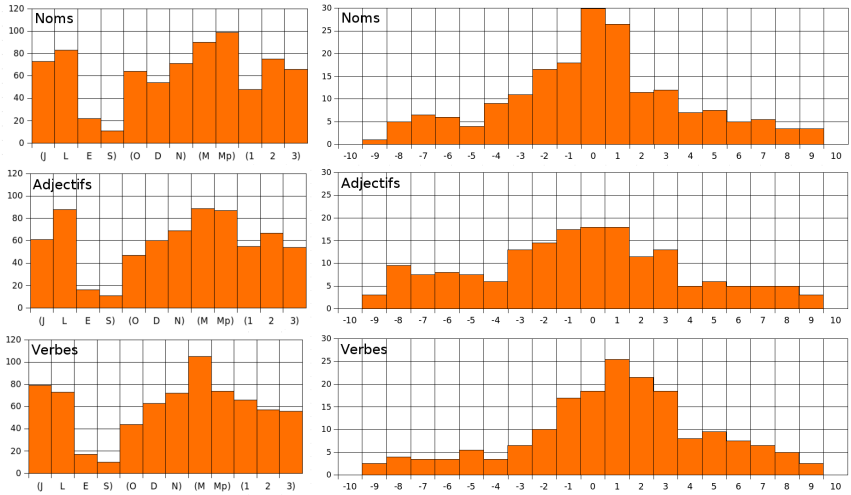


FIG. 1 – Forme et répartition spatiale des indices sélectionnés par l'algorithme génétique appliqué par catégorie grammaticale sur le jeu de 888 indices. Les graphiques de gauche montrent : les proportions d'indices constitués des étiquettes `jeton` (J), `lemme` (L), `ems` (E) ou `smalems` (S) ; les proportions d'indices différenciés par leur position (O), différenciés suivant qu'ils appartiennent au contexte droit ou gauche (D), ou non différenciés (N) ; les proportions d'indices constitués de mots sans distinction (M) ou seulement de mots pleins (Mp) ; et enfin les proportions d'indices constitués d'unigrammes (1), de bigrammes (2) ou de trigrammes (3). Les graphiques de droite montrent où se situent les indices (en prenant la position médiane pour les bigrammes et trigrammes) par rapport au mot à désambiguïser.

Nous avons cherché à en savoir plus sur les indices sélectionnés par l'algorithme génétique appliqué par catégorie grammaticale sur le jeu de 888 indices, c'est-à-dire par la sélection qui obtient les meilleurs résultat et qui correspond à la dernière ligne du tableau 4. La figure 1 résume ces observations pour chacune des catégories grammaticales.

Les graphiques de gauche permettent de remarquer que les étiquettes `jeton` et `lemme` sont bien plus utilisées que les étiquettes `ems` et `smallems` ce qui paraît logique et cohérent avec la littérature. Ils permettent également d'observer que les indices sélectionnés sont constitués en proportions comparables d'unigrammes, de bigrammes et de trigrammes. Cette observation conforte celle que nous avions faite dans (Audibert, 2004), à savoir que les bigrammes[4] et les trigrammes véhiculent une information importante qui ne se retrouve pas dans les unigrammes. Ces graphiques permettent enfin d'observer que la sélection opérée par l'algorithme génétique ne privilégie pas les indices constitués uniquement de mots pleins. Comme nous l'avions remarqué dans (Audibert, 2004), le filtrage consistant à supprimer les mots grammaticaux n'apparaît absolument pas pertinent.

[4] cf. également (Pedersen, 2001a) concernant l'utilisation des bigrammes.

Les graphiques de droite de la figure 1 montrent que la répartition spatiale des indices sélectionnés par l'algorithme génétique diffère suivant la catégorie grammaticale du mot à désambiguïser. Concernant les noms, la répartition des indices est grossièrement symétrique par rapport au mots à désambiguïser et les indices les plus proches sont privilégiés. La répartition des indices pour la désambiguïsation des adjectifs est bien plus aplatie que pour les deux autres catégories grammaticales. De plus, ce sont les adjectifs qui bénéficient le moins de l'amélioration de la précision apportée par la sélection automatique des indices : $3,1pt$ contre $4,4pt$ pour les noms et $5,4pt$ pour les verbes. Comme nous l'avions déjà observé dans (Audibert, 2004), la répartition des indices pour la désambiguïsation des verbes est fortement dissymétrique probablement parce que la désambiguïsation des verbes se fait plus en fonction de leur objet que de leur sujet, la forme sujet-verbe-complément étant la plus fréquente.

5 Conclusion et perspectives

Comme (Mohammad & Pedersen, 2004; Ng & Lee, 2002; Pedersen, 2001a), entre autres, nous pensons que les performance d'un algorithmes de désambiguïsation dépendent principalement de la qualité des indices du contexte considéré plutôt que de la sophistication des algorithmes de désambiguïsation utilisés. Dans cet article, nous avons exposé une expérience consistant à automatiser une sélection d'indices de natures différentes. Ainsi, en réalisant une sélection automatique basée sur un algorithme génétique, nous sommes parvenus à une précision de désambiguïsation, sur les 60 vocables de notre étude, de 79.5%, soit $4,8pt$ de plus que la précision obtenue par le meilleur critère homogène identifié lors de notre étude systématique précédente (Audibert, 2004).

Cette amélioration est importante, mais d'autres espoirs d'améliorations sont à attendre de l'enrichissement des indices disponibles en utilisant, par exemple :
– des indices issus de relations syntaxiques binaires (nom-nom, nom-verbe, adjectif-nom, etc.) ;
– des thésaurus ou des ontologies pour effectuer des généralisations sur les mots du contexte du mot à désambiguïser ;
– des informations sur le thème du texte.

Références

AUDIBERT L. (2003a). Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences. In *10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN-2003)*, p. 35–44, Batz-sur-Mer.

AUDIBERT L. (2003b). *Outils d'exploration de corpus et désambiguïsation lexicale automatique*. PhD thesis, Université de Provence.

AUDIBERT L. (2004). Word sense disambiguation criteria : a systematic study. In *20th International Conference on Computational Linguistics (COLING-2004)*, p. 910–916, Geneva.

BRUCE R., WIEBE J. & PERDERSEN T. (1996). The measure of a model. In E. BRILL & K. W. CHURCH, Eds., *1st Conference on Empirical Methods in Natural Language Processing (EMNLP-1996)*, p. 101–112, Somerset, New Jersey : Association for Computational Linguistics.

CUSSENS J. (1993). Bayes and pseudo-bayes estimates of conditional probability and their reliability. In P. B. BRAZDIL, Ed., *6th European Conference on Machine Learning (ECML-1993)*, p. 136–152, Springer-Verlag, Berlin.

DAELEMANS W., HOSTE V., MEULDER F. D. & NAUDTS B. (2003). Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In *14th European Conference on Machine Learning (ECML-2003)*, Cavtat-Dubrovnik, Croatia.

EDMONDS P. (2002). Introduction to senseval. In *ELRA Newsletter*.

IDE N. & VÉRONIS J. (1998). Word sense disambiguation : The state of the art. In *Computational Linguistics : Special Issue on Word Sense Disambiguation*, volume 24, p. 1–40.

KILGARRIFF A. (1997). Evaluating word sense disambiguation programs : Progress report. In *Speech and Language Technology (SALT-1997) Workshop on Evaluation in Speech and Language Technology*, p. 114–120, Sheffield University, United Kingdom.

MOHAMMAD S. & PEDERSEN T. (2004). Combining lexical and syntactic features for supervised word sense disambiguation. In *Proceedings of CoNLL-2004*, p. 25–32, Boston, MA, USA.

NG H. T. & LEE Y. K. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *7th Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, p. 41–48, Philadelphia, Pennsylvania, USA.

NG H. T. & ZELLE J. (1997). Corpus-based approaches to semantic interpretation in natural language processing. In *Artificial Intelligence Magazine - Special Issue on Natural Language Processing*, volume 18, p. 45–64.

PALMER M. (1998). Are WordNet sense distinctions appropriate for computational lexicons. In *Association for Computational Linguistics Special Interest Group on the Lexicon (ACL-SIGLEX-1998) : SENSEVAL*, Herstmonceux, Sussex, UK.

PEDERSEN T. (2001a). A decision tree of bigrams is an accurate predictor of word sense. In *Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, p. 79 ?–86, Pittsburgh.

PEDERSEN T. (2001b). Machine learning with lexical features : The duluth approach to senseval-2. In *2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, p. 139–142.

REYMOND D. (2001). Dictionnaires distributionnels et étiquetage lexical de corpus. In *5ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL-2002)*, volume 1, p. 479–488, Tours.

VALLI A. & VÉRONIS J. (1999). Etiquetage grammatical de corpus oraux : Problèmes et perpectives. In *Revue Française de Linguistique Appliquée*, volume IV, p. 113–133. Champs-sur-Marne : Association pour le traitement informatique des langues (ASSTRIL).

VÉRONIS J. (1998). A study of polysemy judgements and inter-annotator agreement. In *Programme and Advanced Papers of the Senseval Workshop*, Herstmonceux Castle, England.

VÉRONIS J. (2001). Sense tagging : Does it makes sense. In *Corpus Linguistics*, Lancaster, U.K.

YAROWSKY D. (1994). A comparision of corpus-based techniques for restoring accents in spanish and french text. In *2nd Annual Workshop on Very Large Text Corpora*, p. 19–32, Las Cruces.

Représenter la dynamique énonciative et modale de textes

Delphine BATTISTELLI[1], Marie CHAGNOUX[2]
[1] Lalic – Université Paris IV-Sorbonne, 28 rue Serpente 75006 Paris
[2] France Télécom – Div. R&D, TECH/EASY/Langues Naturelles,
2 avenue Pierre Marzin, 22307 Lannion Cedex
Delphine.Battistelli@paris4.sorbonne.fr,Marie.Chagnoux@free.fr

Résumé. Nous proposons d'exposer ici une méthodologie d'analyse et de représentation d'une des composantes de la structuration des textes, celle liée à la notion de prise en charge énonciative. Nous mettons l'accent sur la structure hiérarchisée des segments textuels qui en résulte ; nous la représentons d'une part sous forme d'arbre et d'autre part sous forme de graphe. Ce dernier permet d'appréhender la dynamique énonciative et modale de textes comme un cheminement qui s'opère entre différents niveaux de discours dans un texte au fur et à mesure de sa lecture syntagmatique.

Abstract. We propose a methodological framework for analyzing and representing the concept of commitment, which is one of the features characterizing textual structure. We emphasize the hierarchical structure of textual segments commitment conveys to. We represent it first as a tree and then as a graph. The latter enables us to access the modal and enunciative textual dynamics, as it shows the path followed through different discursive levels during the syntagmatic reading of a text.

Mots-clés : linguistique textuelle, énonciation, représentation sémantique.

Keywords: textual linguistics, enunciation, semantical representation.

1 Introduction

Notre approche vise à mettre l'accent sur la dynamique d'interprétation d'un texte appréhendée au travers des mécanismes de prise en charge énonciative qui instruisent une structuration de type hiérarchique. Nous proposons d'en donner deux types de représentation : la première met en évidence la structure énonciative et modale du texte, que nous représentons sous la forme d'un arbre ; la deuxième met en évidence la dynamique énonciative et modale du texte, que nous représentons sous la forme d'un graphe. Ces représentations vont toutes deux être convoquées pour deux types d'utilisations que nous visons : la première consiste seulement à rendre possible la construction automatique et la visualisation d'arbres et de graphes à l'écran ; la seconde consiste à utiliser ces représentations comme représentations internes d'un texte pour proposer des parcours entre segments dans le cadre d'une plateforme de navigation textuelle. Nous proposons dans cet article d'expliciter notre démarche. Dans la section 2, nous la situons par rapport aux approches existantes. La section 3 présente les étapes préliminaires de segmentation et d'annotation automatiques qui permettent de construire les représentations selon des modalités détaillées dans la section 4.

2 La prise en compte du contexte énonciatif et modal dans les systèmes de recherche d'information

Les systèmes actuels de recherche d'information tels que les résumeurs automatiques ou les systèmes de questions-réponses ne sélectionnent pas des documents entiers mais des *segments* correspondant à une requête ou à un profil à l'intérieur de ceux-ci. L'extraction de ces segments brise la continuité référentielle à l'œuvre dans un texte (Battistelli et Minel, 2006) : privés de leur contexte, les segments peuvent alors être difficiles à interpréter, en particulier quant à leur ancrage temporel, énonciatif et modal.

2.1 Exemples

L'absence de prise en compte du contexte énonciatif et modal peut entraîner des imprécisions, des contresens ou des conflits comme l'illustrent les trois exemples du tableau 1. Si l'on considère le segment *l'appel à un cessez-le-feu urgent et non pas immédiat constitue un échec*, on infère que l'auteur de l'article dont est extrait ce segment considère que l'appel à un cessez-le-feu est un échec. Or, si on le réinscrit dans sa continuité référentielle de l'extrait 1 du tableau 1, il apparaît que l'auteur ne prend en charge que le principe de citation et qu'un énonciateur second, la Maison-Blanche, réfute cette proposition.

1. La Maison-Blanche s'est, elle, montrée satisfaite des résultats de la réunion de Rome et a réfuté que *l'appel à un cessez-le-feu urgent et non pas immédiat constitue un échec*.
2. 1. Le Hezbollah a capturé deux soldats israéliens qui patrouillaient dans des jeeps blindées *sur le territoire libanais* à la frontière avec Israël. 2. 2. Le mercredi 12 juillet, 8 soldats de Tsahal ont été tués et 2 ont été kidnappés, au cours d'une attaque du Hezbollah *à l'intérieur du territoire israélien*, près de la frontière israélo-libanaise.
3. A propos du lieu de l'enlèvement, les versions diffèrent. Les Israéliens affirment qu'ils ont été capturés près de la ferme collective de Zarit *en territoire israélien* tout près de la frontière libanaise. De son côté, la police libanaise soutient que la capture s'est produite dans la région de Aïta al-Chaab *en territoire libanais* donc, proche de la frontière libano-sraélienne où une unité israélienne avait pénétré le matin même.

Tableau 1 : Extraits de textes portant sur le conflit entre Israël et le Liban durant l'été 2006

L'exemple 2 du même tableau montre que l'extraction de segments isolés de leur contexte énonciatif peut conduire à des conflits. Ainsi les exemples 2.1 et 2.2 sont contradictoires. Les deux extraits réfèrent à la même situation : la capture de deux soldats au Moyen-Orient le 12 juillet 2006. Or, selon l'énonciateur, le lieu de la capture est différent : pour l'association France Palestine, la capture a eu lieu sur le territoire libanais ; pour l'Ambassade d'Israël en France, elle s'est déroulée à l'intérieur du territoire israélien. La reconnaissance de la source peut paraître triviale dans le cas d'application inter-documents puisqu'il suffit d'annoter le texte pour le doter de cette information comme cela est proposé par exemple pour le Web sémantique. Cependant l'examen de l'exemple 3 du tableau 1 montre que le même conflit peut apparaître au sein d'un seul et même texte.

2.2 Annotation *vs.* Représentation

Le texte est la trace d'au moins un acte d'énonciation, cet acte est celui accompli par l'énonciateur principal qui prend en charge l'ensemble du discours. Cependant, certains segments de son discours ne sont pas complètement assumés par cet énonciateur, soit qu'un énonciateur second est convoqué au terme d'une citation, soit que ces segments sont assujettis à un certain degré de plausibilité ou d'intentionnalité. Il devient alors crucial pour les

systèmes de disposer de cette information qui instruit le contexte énonciatif et modal d'un segment textuel, comme en témoignent les trois exemples du tableau 1. Actuellement, de nombreux travaux s'inscrivent dans ce paradigme et proposent des systèmes d'annotation de ce type d'information (Giguet et Lucas 2004 ; Harabagiu et al., 2004 ; Sauri et al., 2005 ; Wilson et Wiebe, 2005). Nous pensons pour notre part que la seule tâche d'annotation des ruptures de prises en charge énonciative n'est pas suffisante et qu'il est nécessaire de les organiser en une structure représentationnelle qui rend compte des liens de type hiérarchique entre les segments annotés. En ayant ainsi repéré et hiérarchisé les différentes prises en charge énonciatives, un traitement informatique est à même de rendre compte de la véritable attribution de tel ou tel segment à tel ou tel énonciateur, que ce soit dans le cadre d'un traitement inter-documents ou intra-document.

2.3 Linguistique textuelle et systèmes de navigation textuelle

Pour des systèmes de navigation textuelle, qui proposent une alternative aux systèmes classiques de résumé automatique et de questions-réponses en permettant une navigation à l'intérieur d'un document parmi des segments annotés inscrits dans leurs co-textes (Bilhaut et al., 2003 ; Crestan et al., 2004 ; Minel et Couto, 2004), disposer de ce type d'informations constitue un enjeu important. En outre, il est intéressant de relever que ces systèmes s'appuient sur les structures discursives identifiées (le plus souvent de type hiérarchique) pour proposer une navigation à l'intérieur des documents ; ce faisant, ils sont directement concernés par les travaux récents de la linguistique textuelle qui a pour objet l'analyse des modes d'organisation discursifs (Péry-Woodley, 2001). Parmi ces modes, la temporalité occupe une place privilégiée très largement étudiée : par l'analyse des indices calendaires en position initiale de phrases qui constituent des marques d'ouverture de cadres temporels (Charolles et Vigier, 2005 ; Le Draoulec et Péry-Woodley, 2005), par l'analyse des variations de prises en charge énonciative et modale de contenus propositionnels (Desclés et Guentchéva, 2000 ; Kronning, 2003), etc.

La méthodologie de la linguistique textuelle consiste à établir le lien entre structures discursives et marques formelles organisées en systèmes. Elle se distingue en cela de celle mise en œuvre dans le cadre de modèles du discours comme ceux de la RST (Mann et Thompson 1988) ou de (Hobbs 1990) qui insistent sur l'indépendance du type de structures qui les intéresse par rapport à toute marque linguistique[1]. Nous situant pour notre part dans une démarche d'analyse empirique des modes de structuration discursive – ici, énonciative et modale –, nous nous inscrivons dans les principes méthodologiques de la linguistique textuelle et donc aussi dans ceux qui fondent selon nous le développement de systèmes de navigation textuelle qui s'appuient directement sur les structures discursives mises en évidence. Nous exposons brièvement dans la section suivante l'étape qui précède celle de la construction de la représentation hiérarchique du texte : la tâche d'annotation. Elle consiste (i) d'une part à spécifier le type d'unités textuelles à annoter et donc à procéder à une segmentation du texte ; (ii) et d'autre part à préciser la typologie des annotations sémantiques à associer à ces unités.

[1] Il est à noter que l'approche de la SDRT (Asher, 1993) est à cet égard quelque peu différente. Il existe en effet depuis plusieurs années une volonté de lier le modèle aux instanciations langagières des relations de discours.

3 Segmentation et annotation des segments textuels

Nous présentons dans cette section les principes qui régissent l'annotation et en gouvernent l'automatisation. Au niveau discursif, le texte est considéré comme une suite de propositions en relations de rupture ou de continuité. Ces relations qui conduisent à considérer des segments homogènes regroupant plusieurs propositions sont définies lors de la construction de la représentation. Au niveau de l'annotation, le texte est segmenté en propositions par le système PROPOS (Wonsever, 2004) avant d'être annoté. Les critères d'annotation sont fondés sur les types de décrochage de prise en charge et sur la typologie des référentiels proposée par (Desclés, 1995) et précisée par (Chagnoux, 2006). Ces critères permettent d'inscrire les propositions sur différents référentiels :

- Le référentiel énonciatif *RE*, sur lequel le contenu propositionnel est complètement assumé par un énonciateur *E* ; il peut être global (noté *REG*) quand il s'agit de l'énonciateur premier ou local (noté *REL*) quand il s'agit d'un discours rapporté, direct ou indirect, ou médiatisé ;
- Le référentiel possible *RP*, sur lequel le contenu propositionnel est considéré comme possible éventuel (dans ce cas, le référentiel est noté *RPE*) ou possible contre-factuel (ici, le référentiel est noté *RPC*) ;
- Le référentiel mental noté *RM*, où s'inscrivent des propositions introduites par *il pense que*, *il croit que*, etc., à rapprocher des méta-représentations décrites par (Recanati, 2000).

L'objectif de cet article étant davantage de s'intéresser aux mécanismes de construction des structures hiérarchiques qui résultent de l'identification de ces différents référentiels dans les textes, nous ne présentons pas ici les analyses linguistiques détaillées sur lesquelles se fondent cette typologie. On pourra se reporter à (Guentchéva, 1996), (Desclés et Guentchéva, 2000) ou (Chagnoux, 2006) pour une présentation des analyses de discours directs, indirects, indirects libres et de la médiation.

L'identification et l'annotation des segments s'appuient sur la présence de marqueurs linguistiques de différentes natures : temps et modes des verbes conjugués, déictiques, constructions syntaxiques, lexèmes, adverbes, etc. Nous insistons sur le fait que ces catégories fonctionnent en systèmes et que c'est l'expression de patrons complexes qui rend compte des opérations de rupture ou au contraire de cohésion. Une base de marqueurs et une base de règles heuristiques ont été implémentées dans Chronotext[2] afin d'annoter les référentiels. Le tableau 2 détaille la segmentation et l'annotation d'un extrait court : l'extrait 1 est segmenté selon le format proposé en 2 par PROPOS (résumé en 3 pour une meilleure lisibilité) et annoté comme dans 4 par Chronotext.

Les annotations de 4 sont obtenues par les règles présentées dans le tableau 3 et qui peuvent être glosées comme suit : nous partons du postulat que tout texte s'inscrit dans un référentiel énonciatif global par défaut. Tant qu'aucun marqueur ne permet de les assigner à un autre référentiel, toutes les propositions relèvent de ce référentiel énonciatif global. Ici, c'est le cas de P1.

[2] *Chronotext*, présenté en détail dans (Chagnoux 2006) permet la reconnaissance des différents référentiels grâce à une base de 76 règles et de 708 marqueurs organisé dans 119 classes implémentés dans *Semantext*, plate-forme de collecte et d'organisation de connaissances linguistiques pour l'annotation des textes.

1. Ses cadres affirment qu'ils ont les moyens de procéder à une escalade de grande envergure si Tsahal poursuit ses raids sur le sol libanais.
2. <PHRASE_TEXTE><PROP COD=pli1 NIVEAU=1 INDEP VERBE = affirment>Ses cadres affirment </PROP><PROP COD=prl1 NIVEAU=2 INSER IP=mot (que) VERBE = ont>qu'ils ont les moyens de procéder à une escalade de grande envergure ><PROP COD=gf21 NIVEAU=3 INDEP VERBE = poursuit>si Tsahal poursuit ses raids sur le sol libanais. </PROP></PROP></PHRASE_TEXTE
3. [P1 : Ses cadres affirment] [P2 : qu'ils ont les moyens de procéder à une escalade de grande envergure] [P3 : si Tsahal poursuit ses raids sur le sol libanais.]
4. <REG> Ses cadres affirment </REG> <REL> qu'ils ont les moyens de procéder à une escalade de grande envergure </REL> <RPE> si Tsahal poursuit ses raids sur le sol libanais.< /RPE>

Tableau 2 : Segmentation et annotation d'un extrait de texte

En revanche, la présence du patron *&ouverture_citation + que*[3] permet d'identifier automatiquement que P2 relève d'un référentiel énonciatif local. L'annotation de P3 résulte de l'introduction de la proposition courante par *si* associée à la règle de concordance des temps. . Des règles de récursivité définissent des conditions de propagation qui permettent à l'annotation sémantique d'une proposition d'être associée à la proposition suivante afin d'annoter l'ensemble des propositions du texte.

1. NOT ($proposition$ contains (/$Refentiel-possible/OR /$Referentiel-mental/ OR/Referentiel-enonciatif2/) AND($proposition$ contains(/&deictique/ OR /&interjection/)) AND NOT ($proposition$contains (/&Nactualise/ OR /&PS/)
2. (($proposition justbefore $proposition$)contains /&ouverture-citation/) AND ($proposition$ contains/ &que/)
3. ($proposition$ starts-with /&si/) AND ($proposition$ contains /&present/) AND ($proposition justbefore $proposition$ contains /&present/)

Tableau 3 : Exemples de règles d'annotation des segments

4 Représentation de la sémantique énonciative et modale d'un texte

La tâche de construction de la représentation oblige à spécifier les relations entre les unités distinguées lors de l'étape précédente. Le fait de considérer une structure textuelle hiérarchisée est classique en analyse du discours (Mann et Thompson, 1988 ; Hobbs, 1990, Asher 1993). Comme le souligne (Cornish 2006), le « scripteur » réalise linguistiquement une organisation qui passe par la constitution de segments, la hiérarchisation de ces segments et leur mise en relation. Autrement dit, un texte est vu comme un ensemble de segments qui entrent en relation d'inclusion ou de succession[4] et il peut donc être représenté à l'aide d'une structure d'arbre. Nous proposons d'exposer ici les principes de construction de la représentation « statique » sous forme d'arbre de la structure de l'extrait (section 4.1) ; puis d'exposer ceux relatifs à la construction de la représentation sous forme de graphe de la « dynamique » du texte (section 4.2). Cette dernière permet de visualiser le cheminement qui s'opère entre différents niveaux de discours dans un texte au fur et à mesure de sa lecture

[3] Dans *Chronotext*, *&ouverture_citation* est le nom de la classe contenant plus de 90 verbes susceptibles d'introduire les propos d'un tiers comme *s'exclamer*, *dire*, *hurler*, etc.

[4] Ces relations peuvent éventuellement faire l'objet d'une typologie sémantique comme c'est le cas pour les relations dites de discours (narration, élaboration, …) proposées par exemple par (Mann et Thompson 1998) ; il reste qu'elles renvoient toutes à une structuration arborescente du texte.

syntagmatique. Ces deux représentations sont construites à partir des segments annotés du texte présenté dans le tableau 4.

Hier après-midi, quarante-huit heures après l'ouverture du front libano-israélien, le Hezbollah qui, selon son chef Hassan Nasrallah, disposerait de plus de 10 000 roquettes et missiles pointés sur Israël, avait déjà frappé 165 fois le territoire de l'Etat hébreu. Et ce ne pourrait être qu'un début. Ses cadres affirment qu'ils ont les moyens de procéder à une escalade de grande envergure si Tsahal poursuit ses raids sur le sol libanais.

Tableau 4 : *Incipit* de l'article "Les chiites fous du Parti de Dieu", *Libération*, (15/072006)

4.1 Représenter la structure énonciative et modale sous forme d'arbre

Pour faire émerger la structure textuelle, l'annotation n'est plus seulement faite au niveau propositionnel par des balises XML mais se réalise au niveau segmental via un modèle de représentation du texte défini par (Couto 2006) et spécifié dans une DTD. Ce modèle permet de définir des Unités Textuelles (UT) à partir de propositions connexes mais surtout de traiter le problème de la discontinuité par une opération de Séquence qui rétablit la cohésion entre des segments discontinus dans le texte. A chaque UT est associé un attribut Nature qui définit la nature du segment. Chaque type de référentiel convoque des attributs différents qui permettent d'encoder toutes les propriétés qui lui sont associées. La figure 1 présente la mise au format du texte dont est extrait l'exemple 1 du tableau 2. La Tête contient les opérations de Séquence et le Corps contient les différentes propositions du texte.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE Texte SYSTEM "DocumentNaviTexte.dtd">
<Texte>
   <Tete>
      <Sequence Type="Regroupement" Nro="1" >
         <UTP Type=" Segment_referentiel " Nro="1"/>
         <UTP Type=" Segment_referentiel " Nro="4"/>
      </Sequence>
   </Tete>
   <Corps>
      <UT Type=" Segment_referentiel " Nro="1">
            <Attribut Nom="Nature">Enonciatif_global</Attribut>
      <UT Type=" Proposition" Nro="1">
            <Chaine>Hier après-midi, quarante huit heures après l'ouverture du front libano-israélien, le
Hezbollah avait frappé 165 fois le territoire de l'Etat hébreu</Chaine></UT>
      <UT Type=" Segment_referentiel " Nro="2">
            <Attribut Nom="Nature">Enonciatif_local
</Attribut>
         <UT Type=" Proposition " Nro="2">
            <Chaine> qui, selon son chef Hassan Nasrallah, disposerait de plus de 10 000 roquettes et
missiles pointés sur Israël, </Chaine></UT></UT>
         <UT Type=" Segment_referentiel " Nro="3">
            <Attribut Nom="Nature">Possible_eventuel</Attribut>
         <UT Type=" Proposition " Nro="3">
            <Chaine> Et ce pourrait n'être qu'un début </Chaine></UT></UT>
         <UT Type=" Segment_referentiel " Nro="4">
            <Attribut Nom="Nature">Enonciatif_global</Attribut>
         <UT Type=" Proposition " Nro="4">
            <Chaine> Ses cadres affirment</Chaine></UT></UT>
         <UT Type=" Segment_referentiel " Nro="5">
            <Attribut Nom="Nature">Enonciatif_local</Attribut>
         <UT Type=" Proposition " Nro="5">
            <Chaine>  qu'ils  ont  les  moyens  de  procéder  à  une  escalade  de  grande  envergure
</Chaine></UT></UT>
         <UT Type=" Segment_referentiel " Nro="6">
            <Attribut Nom="Nature">Enonciatif_local</Attribut>
         <UT Type=" Proposition " Nro="6">
            <Chaine> si Tsahal poursuit ses raids sur le sol libanais </Chaine></UT></UT></UT></UT>
   </Corps>
   </Texte>
```

Figure 1. Texte annoté en conformité avec la DTD définie dans (Couto 2006)

A partir de cette organisation des segments, un arbre est généré via Graphviz[5]. La figure 2 présente le graphe associé à cet extrait. La racine de l'arbre correspond à *REG* ; les nœuds correspondent aux différents référentiels sur lesquels s'inscrivent les propositions du texte.

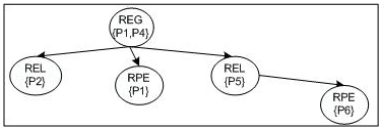
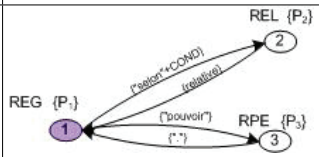
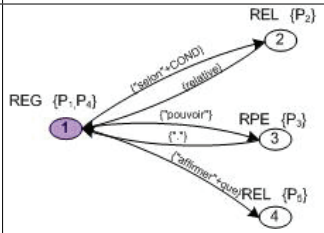


Figure 2 : Arbre de la structure énonciative et modale de l'extrait 2

4.2 Représenter la dynamique énonciative et modale sous forme de graphe

Le phénomène des différentes prises en charge énonciative à l'œuvre dans un texte peut être rendu par une structure de *graphe*. Dans ce cas, on met l'accent sur la *dynamique* énonciative interne du texte comme l'illustre le tableau 5 où se construit progressivement le graphe associé au texte du tableau 3.

Toute énonciation ouvre un référentiel énonciatif global (REG) : on construit ici REG, premier nœud du graphe sur lequel on situe P_1	
Bien que situé sur REG, P_1 contient des indices de rupture énonciative (*selon* suivi d'un emploi du conditionnel) qui ouvrent un référentiel énonciatif local REL	
C'est sur ce référentiel que s'inscrit la proposition P_2. La fin de la proposition relative permet de fermer le référentiel REL et de revenir sur REG. P_3 contient un indice de rupture modale (*pouvoir*) qui ouvre un référentiel du possible éventuel RPE où situer P_3. Il se referme grâce à la présence du signe de ponctuation ".	
P_4 contient un indice de rupture énonciative (*affirmer que*) qui ouvre un référentiel énonciatif local REL où se situe P_5.	

[5] GraphViz est un logiciel libre qui permet de générer automatiquement des arbres et des graphes.

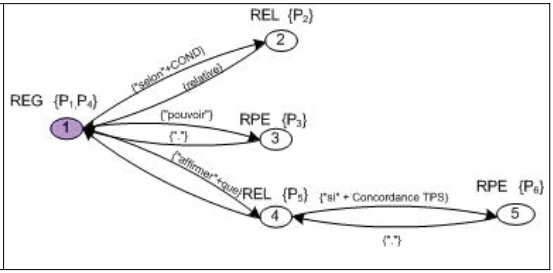
P_6 contient un indice de rupture modale (*si*) qui ouvre un référentiel du possible éventuel RPE où situer P_6. Il se referme grâce à la présence du signe de ponctuation ". Enfin, la fin du texte avec un point, en tant que fin de l'énonciation, ferme tous les référentiels ouverts et le graphe s'achève donc par un retour à REG.	

Tableau 5 : Etapes de construction du graphe de la dynamique énonciative et modale du texte

Les nœuds correspondent toujours aux référentiels identifiés. Les arcs indiquent les parcours entre les différents niveaux de discours et sont étiquetés avec les marqueurs qui font passer d'un niveau à un autre. Il faut souligner que cette structure encode la structure arborescente sous-jacente puisque les bi-arcs entre nœuds correspondent exactement aux liens de l'arbre. Lors de la construction de la structure textuelle, la principale difficulté réside dans l'identification automatique de la fermeture des segments repérés. Cette difficulté se pose également dans les autres modèles qui rendent compte de la structure textuelle (Charolles et Vigier 2005), diverses stratégies sont mises en œuvre pour résoudre ce point comme par exemple dans cet extrait le retour systématique au premier nœud du graphe en fin de texte.

5 Conclusion

Dans cet article, nous avons proposé un modèle de représentation de la dynamique énonciative et modale ainsi que son automatisation via la chaîne de traitement que synthétise la figure 3.

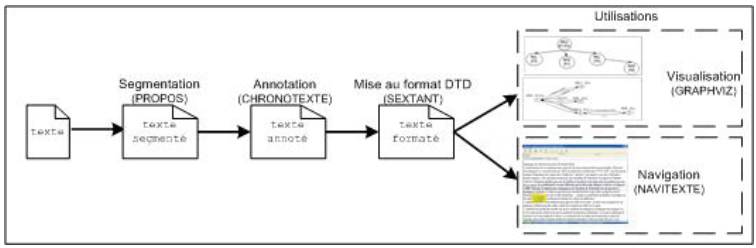


Figure 3: Chaîne de traitement pour l'automatisation de la génération de graphes

Cette chaîne est actuellement en cours de réalisation et est testée sur un corpus de 25 textes (5 textes journalistiques, 5 textes tirés de la Wikipedia, 5 incipit de roman, 5 textes scientifiques et 5 textes "entreprises"). Les tâches de segmentation et d'annotation sont opérationnelles et nous travaillons actuellement d'une part à la mise au format automatique dans la DTD choisie et d'autre part à la génération automatique de l'arbre et du graphe via Graphviz. Il nous faut encore résoudre un certain nombre de difficultés : lors de l'annotation automatique des propositions, il est parfois difficile de distinguer automatiquement la nature de tous les référentiels. Ainsi par exemple, le repérage des assertions introduites par "selon" ou" pour" en position antéposée produit encore trop de bruits en identifiant des propositions comme "selon

l'appartenance sociale" ou "pour le développement". Les règles qui traitent les phénomènes modaux doivent encore être affinées pour permettre une meilleure annotation. Il faudrait également à terme pouvoir regrouper les segments discontinus pris en charge par un même énonciateur.
En ce qui concerne l'évaluation, comme l'explicite (Chaudiron, 2004), la difficulté qu'il y a à construire des protocoles d'évaluation est particulièrement accrue dans le domaine sémantique. A défaut de pouvoir évaluer la pertinence de la représentation sémantique proposée ici, nous pensons qu'illustrer son utilisation pour d'autres tâches permet de valider notre approche. Les premiers résultats que nous avons obtenus montrent d'une part qu'il nous est possible de traiter tout type de texte et, par là même, de proposer une typologie basée sur les mécanismes énonciatifs (Battistelli et Chagnoux, 2006). D'autre part, le format d'encodage des informations permet un certain nombre d'opérations de navigation à l'intérieur d'un texte dans une plate-forme comme celle de (Couto et Minel, 2006). Notre approche se distingue de la « simple » approche d'annotation des ruptures de prise en charge énonciative et modale puisque nous proposons de construire, à partir des textes annotés, des représentations des structures discursives. Notre modèle de représentation discursive prend non seulement en compte les ruptures de prises en charge énonciative mais les organise en une structure hiérarchique, ce qui constitue un enjeu fondamental pour les systèmes actuels de recherche d'information. Par ailleurs, ce projet prend toute sa légitimité dans le cadre de plateformes de navigation textuelle qui s'appuient directement sur ces structures pour proposer des parcours de navigation entre segments textuels.

Références

ASHER N. (1993). *Reference to Abstract objects in Discourse.* Kluwer Academic Publishers.

BATTISTELLI D., CHAGNOUX M. (2006). Vers une typologie de mécanismes discursifs. *Actes de la journée ATALA "Typologie de textes"*, Paris.

BATTISTELLI D., MINEL J.-L. (2006). Les systèmes de résumé automatique : comment assurer une continuité référentielle dans la lecture des textes, in G. Sabah (Ed.), *Compréhension des langues et interaction*, 295-330.

BATTISTELLI D., MINEL J.-L., SCHWER S. (2006). Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies. *TAL* vol. 47/2.

BILHAUT F., HO-DAC M., BORILLO A., CHARNOIS T., ENJALBERT P., LE DRAOULEC A., MATHET Y., MIGUET H., PÉRY-WOODLEY M.-P., SARDA L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique, Actes *TALN'03*, Batz-sur-Mer, 315-320.

CHAGNOUX M. (2006). Temporalité et aspectualité dans les textes français. Modélisation sémantico-cognitive et traitement informatique. *Thèse de l'Université Paris-Sorbonne*.

CHAROLLES M., VIGIER D. (2005). Les adverbiaux en position préverbale : portée cadrative et organisation des discours, *Langue Française*, 148, 9-30.

CHAUDIRON, S. (2004). *L'évaluation des systèmes de traitement de l'information*. Lavoisier.

COUTO J. (2006). Modélisation des connaissances pour une navigation textuelle assistée. La plate-forme logicielle NaviTexte. *Thèse de l'Université Paris-Sorbonne*.

COUTO J., MINEL J.-L. (2006). SEXTANT, un langage de modélisation des connaissances pour la navigation textuelle. Actes *ISDD'06*, Caen, 80-90.

CORNISH F. (2006). Relations de cohérence en discours : critères de reconnaissance, caractérisation et articulation cohésion-cohérence, Journée d'étude du CRISCO *Organisation des textes et cohérence des discours*, Université de Caen, texte publié en ligne en 2006

CRESTAN É., DE LOUPY C., MANIGOT L. (2004). Analyses sémantiques pour la navigation textuelle ; Actes *CIDE 7*, La Rochelle ; 22-25 juin 2004, 293-308.

DESCLES J.-P. (1995). Les référentiels temporels pour le temps linguistique. *Modèles linguistiques* 16, 9-36.

DESCLÉS J.-P., GUENTCHEVA Z. (2000). Enonciateur, locuteur, médiateur. Erikson Ph et Monod-Becquelin A (éds). *Les rituels du dialogue.* Nanterre : Société d'éthnologie.

GUENTCHEVA Z. (éd.) (2000). *L'énonciation médiatisée*. Louvain-Paris : Peeters.

GIGUET E., LUCAS N. (2004). La détection automatique des citations et des locuteurs dans les textes informatifs. *Le Discours rapporté dans tous ses états : question de frontières,* Paris, L'Harmattan.

HARABAGIU M., MAIORANO S.J., MOSCHITTI A., BEJAN A.C. (2004). Intentions, Implicatures and Processing of Complex Questions. Actes *HLT-NAACL Workshop on Pragmatics of Q. A.*

HOBBS J.R. (1990). Chap. 5 : The coherence and structure of discourse, *Literature and Cognition*, Leland Standford Junior University, Calif: *CSLI Lecture Notes* 21, 83-114.

KRONNING H. (2003). Modalité et évidentialité. Birkelund, M., Boysen, G. & Kjærsgaard, P. S. (éds). *Aspects de la Modalité*, *Linguistische Arbeite*n 469, 131-151.

LE DRAOULEC A., PÉRY-WOODLEY M.-P. (2005). Encadrement temporel et relations de discours, *Langue Française* 148, 45-60.

MANN W.C., THOMPSON S.A. (1988). Rhetorical Structure Theory: toward a functional theory of text organization*, Text* 8(3), 243-281.

PÉRY-WOODLEY M.-P. (2001). Présentation du numéro : Cohérence et relations de discours à l'écrit, *Verbum,* XXIII, 1.

RECANATI F. (2000). *Oratio Obliqua, Oratio Recta : An Essay on Metarepresentation*. Cambrige : MIT Press.

SAURI R., VERHAGEN M., PUSTEJOVSKY J. (2005). Annotating and recognizing Event Modality in Text. Actes *FLAIRS'06,* Melbourne Beach, Florida.

WILSON T., WIEBE J. (2005). Annotating Attributions and Private States, *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor, Michigan.

WONSEVER D. (2004). Repérage automatique des propositions par exploration contextuelle*. Thèse de l'Université Paris-Sorbonne.*

Segmentation en super-chunks

Olivier BLANC, Matthieu CONSTANT, Patrick WATRIN
IGM, Université de Marne-la-Vallée & CNRS
{oblanc,mconstan,watrin}@univ-mlv.fr

Résumé. Depuis l'analyseur développé par Harris à la fin des années 50, les unités polylexicales ont peu à peu été intégrées aux analyseurs syntaxiques. Cependant, pour la plupart, elles sont encore restreintes aux mots composés qui sont plus stables et moins nombreux. Toutefois, la langue est remplie d'expressions semi-figées qui forment également des unités sémantiques : les expressions adverbiales et les collocations. De même que pour les mots composés traditionnels, l'identification de ces structures limite la complexité combinatoire induite par l'ambiguïté lexicale. Dans cet article, nous détaillons une expérience qui intègre ces notions dans un processus de segmentation en super-chunks, préalable à l'analyse syntaxique. Nous montrons que notre chunker, développé pour le français, atteint une précision et un rappel de 92,9 % et 98,7 %, respectivement. Par ailleurs, les unités polylexicales réalisent 36,6 % des attachements internes aux constituants nominaux et prépositionnels.

Abstract. Since Harris' parser in the late 50's, multiword units have been progressively integrated in parsers. Nevertheless, in the most part, they are still restricted to compound words, that are more stable and less numerous. Actually, language is full of semi-frozen expressions that also form basic semantic units : semi-frozen adverbial expressions (*e.g.* time), collocations. Like compounds, the identification of these structures limits the combinatorial complexity induced by lexical ambiguity. In this paper, we detail an experiment that largely integrates these notions in a procedure of segmentation into super-chunks, preliminary to a parser. We show that the chunker, developped for French, reaches 92.9% precision and 98.7% recall. Moreover, multiword units realize 36.6% of the attachments within nominal and prepositional phrases.

Mots-clés : chunker, super-chunks, analyse syntaxique, patrons lexico-syntaxiques.

Keywords: chunker, super-chunks, syntactic analysis, lexico-syntactic patterns.

1 Introduction

Depuis l'analyseur syntaxique élaboré par l'équipe d'Harris à la fin des années 50 (Joshi & Hopely, 1997), les unités polylexicales ont progressivement été intégrées au processus d'analyse (Nivre & Nilsson, 2004). Cependant, dans la plupart des cas, elles sont restreintes aux mots composés, plus stables et moins nombreux. La langue regorge pourtant d'expressions moins figées qui peuvent également être considérées comme des unités sémantiques de base : les expressions adverbiales semi-figées et les collocations. De même que pour les composés, l'identification de ces structures facilite l'analyse syntaxique en limitant considérablement la combinatoire induite par l'ambiguïté lexicale.

Pour étudier ce phénomène, nous avons implémenté un *chunker*[1] reposant sur la notion de *super-chunk*. Ces structures diffèrent de la notion communément associée aux chunks (Abney, 1996; Karlsson *et al.*, 1995; Federici *et al.*, 1996; Ait-Mokhtar & Chanod, 1997) en ce qu'elles peuvent intégrer des attachements adjectivaux et/ou prépositionnels. Le super-chunk est donc une unité non récursive qui s'arrête à un élément lexical des classes N, V, A, Adv, ou à un élément complexe (MWU) appartenant à ces mêmes classes. Ainsi, par exemple, les séquences *chiffres d'affaires brut* et *marge d'exploitation*, étiquetées `N` (nom) lors de l'analyse lexicale, seront traîtées comme des mots simples durant la phase de segmentation[2]. Dans ce cas, la réduction de l'ambiguïté est évidente. Appréhendée de manière compositionnelle, la séquence *chiffres d'affaires brut* conduit à 24 analyses que nous linéarisons complètement si l'on envisage la collocation dans son ensemble. Par ailleurs, cette seule entrée lexicale nous permet de résoudre un double attachement (prépositionnel et adjectival), facilitant ainsi l'indentification des constituants.

Notre chunker s'inscrit dans un projet plus large visant l'analyse syntaxique du français. Telle que nous la concevons, cette analyse opère en trois phases de rafinement successifs : (1) la segmentation lexicale du texte en unités simples et complexes ; (2) la reconnaissance et l'étiquetage des super-chunks ; (3) l'attachement en constituants. Une illustration de cette procédure incrémentale est donnée au sein du tableau 1. Dans cet exposé, nous ne détaillerons pas plus en profondeur les caractéristiques de l'analyseur et limiterons notre propos à la segmentation en super-chunks. Nous nous concentrerons tout d'abord sur le module de segmentation lexicale en présentant les ressources utilisées. Nous montrerons comment une partie d'entre elles a été apprise automatiquement et comment nous les appliquons aux textes. Nous décrirons ensuite le module de segmentation en super-chunks inspiré par (Abney, 1996), et détaillerons la procédure de désambiguïsation. Enfin, nous évaluerons les performances de notre chunker et montrerons son intérêt pour la résolution d'attachements lexicaux.

NIVEAU	EXEMPLE
`Text`	Le groupe de télécommunications néerlandais KPN a annoncée avoir acquis une participation de 77,5 % dans le troisième opérateur allemand de téléphonie mobile E-Plus.
`Lexique`	Le [$_N$ groupe de télécommunications] néerlandais KPN a annoncé avoir acquis une participation de 77,5 % dans le troisième [$_N$ opérateur allemand de téléphonie mobile] E-Plus.
`Super-Chunk`	Le [$_N$ groupe de télécommunications] [$_{XA}$ néerlandais] KPN a annoncé [$_{XVI}$ avoir acquis] une participation de 77,5 % dans le [$_{XA}$ troisième] [$_N$ opérateur allemand de téléphonie mobile] E-Plus.
	[$_{XN}$ Le groupe de télécommunications] [$_{XA}$ néerlandais] [$_{XN}$ KPN] a annoncé [$_{XVI}$ avoir acquis] [$_{XN}$ une participation] de [$_{XN}$ 77,5 %] dans [$_{XN}$ le troisième opérateur allemand de téléphonie mobile E-Plus].
	[$_{XN}$ Le groupe de télécommunications] [$_{XA}$ néerlandais] [$_{XN}$ KPN] [$_{XV}$ a annoncé avoir acquis] [$_{XN}$ une participation] [$_{XP}$ de 77,5 %] [$_{XP}$ dans le troisième opérateur allemand de téléphonie mobile E-Plus].
`Syntagme`	[$_{N0}$ Le groupe de télécommunications néerlandais KPN] [$_V$ a annoncé avoir acquis] [$_{N1}$ une participation de 77,5 % dans le troisième opérateur allemand de téléphonie mobile E-Plus].

TAB. 1 – Processus global

[1]Les développements informatiques présentés dans ce travail reposent, en grande partie, sur la plate-forme logicielle Outilex (Blanc & Constant, 2006), développée à l'Université de Marne-la-Vallée (IGM).

[2]Notons que les informations morpho-syntaxiques sont héritées de la tête lexicale de l'unité complexe (*i.e. marge* et *chiffre*). De plus, nous associons à ces informations la structure interne de l'unité complexe (*i.e. nom-préposition-nom-adjectif* et *nom-préposition-nom*) afin de permettre une éventuelle décompression du tout (dans le but d'un étiquetage, par exemple).

2 Segmentation lexicale

Le processus de segmentation lexicale constitue la part fondamentale de notre chunker. Nous détaillons, dans cette section, les ressources utilisées de même que leur mode d'application.

Les ressources lexicales responsables de la segmentation se présentent sous deux formes : un ensemble de dictionnaires morpho-syntaxiques et une bibliothèque de grammaires locales. Ces ressources sont soit développées manuellement soit acquises automatiquement à partir de textes bruts.

2.1 Ressources lexicales construites manuellement

Les ressources lexicales développées manuellement s'organisent en un dictionnaire de formes fléchies (Courtois, 1990; Courtois *et al.*, 1997) et un réseau de 190 graphes ou grammaires locales[3].

Le dictionnaire compte 746 198 formes simples et 249 929 formes complexes (dont 245 436 noms[4]). Chaque entrée lexicale s'organise autour d'une forme fléchie, d'un lemme, d'une partie du discours, d'informations morphologiques (*e.g.* genre et nombre), d'informations syntaxiques (*e.g.* la structure interne des mots composés) et d'informations sémantiques (*e.g.* trait humain).

La bibliothèque de grammaires locales lexicalisées décrit un ensemble d'unités polylexicales[5]. Un exemple de grammaire locale est donnée à la figure 1. Cette grammaire décrit des adverbes de date et reconnaît des séquences comme *en mars 2007* et *cinq minutes plus tard*. Les chaînes entre < et > définissent des masques lexicaux[6] (*i.e.* les symboles terminaux). <*minute*>, par exemple, désigne les formes fléchies dont le lemme est *minute* (*i.e. minute* et *minutes*). Les sommets grisés sont, quant à eux, des références à d'autres graphes (*i.e.* les symboles non terminaux).

Notons que le graphe de la figure 1 définit un *transducteur* dont la sortie permet le balisage des séquences reconnues. Chaque adverbe de temps décrit par cette grammaire sera dès lors augmenté de l'étiquette `ADV+time`.

2.2 Collocations nominales et apprentissage

Outre les ressources lexicales développées manuellement, notre analyseur lexical intègre un ensemble de collocations nominales (*i.e.* des séquences de mots qui cooccurrent plus souvent qu'à la normale) apprises automatiquement. De cette manière, nous souhaitons favoriser la modularité de notre approche afin de la rendre viable dans un contexte applicatif réel tel que l'extraction d'information.

[3]Les grammaires locales sont des réseaux de transitions récursifs représentés sous la forme de graphes reconnaissant des langages algébriques (Gross, 1997; Woods, 1970). Elles permettent une représentation aisée des contraintes lexico-syntaxiques dans un contexte local.

[4]En marge des noms (*e.g. pomme de terre*, *faux témoignage*), il contient un ensemble de prépositions (*e.g. au milieu de*, *à cause de*), d'adverbes (*e.g. par ailleurs*, *en pratique*) et de conjonctions (*e.g. bien que*, *pendant que*)

[5]Des noms (*e.g. ministre anglais de l'Agriculture*), des prépositions (*e.g. à dix kilomètres au nord de*), des déterminants numériques (*e.g. vingt-sept*) et nominaux (*e.g. dix grammes de*) et des adverbes (*e.g. en octobre 2006*)

[6]Un masque lexical est une entrée lexicale sous-spécifiée équivalente à une structure de traits.

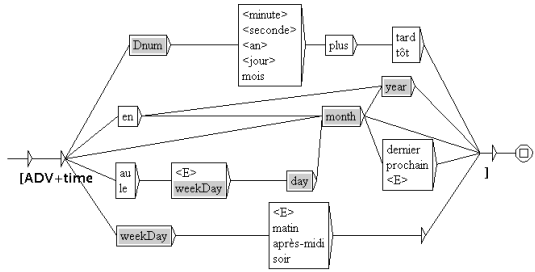


FIG. 1 – Une grammaire locale d'adverbes de date

Pour extraire les collocations, nous avons appliqué la méthode développée dans (Watrin, 2006) à un corpus de dépêches journalistiques d'un million de mots. Cette méthode, inspirée par (Daille, 1995), opère en trois étapes. Dans un premier temps, le corpus d'apprentissage est étiqueté, afin d'évacuer toute ambiguïté (principale source de bruit en extraction) et lemmatisé, pour permettre la généralisation des résultats. Ensuite, un ensemble de patrons syntaxiques, formalisant les structures de collocations, est appliqué au texte afin d'extraire les candidats termes. Finalement, les séquences identifiées sont évaluées statistiquement à l'aide du *log-likelihood* : (Dunning, 1993), pour les bigrammes et (Seretan *et al.*, 2003), pour les trigrammes.

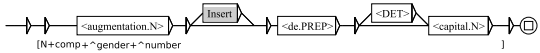


FIG. 2 – Collocation : *augmentation de capital*

Le processus d'extraction associe à chaque collocation sa structure interne. Cette structure nous permet de générer automatiquement les grammaires locales qui seront utilisées par le module de segmentation lexicale. Notons que ces grammaires locales prennent en compte d'éventuels modifieurs. Ainsi, par exemple, la grammaire associée à la collocation *augmentation de capital* (*cf.* FIG. 2) reconnaîtra la séquence *augmentations exceptionnelles de capital*.

L'extraction menée dans le cadre de cette expérience nous a permis d'isoler 1 953 formes canoniques (1 330 bigrammes et 163 trigrammes). Le nombre de collocations extraites pourrait paraître léger mais se justifie pleinement. Nous souhaitons automatiser au maximum le processus d'apprentissage tout en minimisant autant que possible le taux d'erreur. Par conséquent, nous utilisons des contraintes statistiques très fortes qui, si elles réduisent considérablement le nombre de collocation, assure la pertinence des résultats.

D'un point de vue pratique, nous avons observé que 69,1 % des bigrammes et 86,5 % des trigrammes extraits présentent une structure en *préposition-nom*. Ce constat appuie, selon nous, notre hypothèse d'un attachement au niveau lexical et justifie le repérage et l'étiquetage des collocations.

2.3 Application des ressources lexicales

Le module de segmentation lexicale se divise en deux étapes : (1) consultation du dictionnaire et (2) application des grammaires locales lexicalisées. Le programme de consultation du diction-

naire permet d'associer à chaque token toutes les étiquettes linguistiques potentielles et permet également de reconnaître et étiqueter les mots composés. La sortie de ce processus est un automate acyclique dans lequel chaque transition correspond à une entrée lexicale. De cette manière, nous pouvons conserver la totalité de l'ambiguïté. Les grammaires locales sont ensuite appliquées à cet automate, qui est alors augmenté des étiquettes associées aux unités polylexicales identifiées.

Bien que nous cherchions à conserver l'ambiguïté le plus loin possible dans notre chaîne de traitement, notre analyseur permet d'éviter certaines ambiguïtés *artificielles* en supprimant les analyses très rares du dictionnaire. Ainsi, par exemple, les analyses de *a* et *par* comme nom sont enlevées. Pour éviter le silence que peut provoquer la suppression de ces analyses, nous recourrons à un jeu des grammaires locales spécialisées formalisant de manière très précises leurs contextes d'apparition. Dès lors, la forme *par* sera toujours étiquetée *préposition*, sauf dans le cas où elle se trouve dans un contexte lexical particulier tel que *16 au-dessous du par* ou *faire le par*. Dans ce cas, elle sera également analysée comme nom.

3 Segmentation en super-chunks

La segmentation en super-chunks est également incrémentale. Elle consiste en une cascade de transducteurs appliqués à l'automate du texte. L'automate est ainsi augmenté à chaque étape des super-chunks identifiés. La cascade comporte huit étapes et utilise un réseau de 18 graphes reconnaissant successivement :

- les chunks adverbiaux (`XADV`) : les suites d'adverbes simples et les expressions adverbiales reconnues durant l'analyse lexicale ;
- les chunks adjectivaux (`XA`) : les suites d'adjectifs simples pouvant être précédées par un adverbe ;
- les chunks nominaux (`XN`) : les groupes nominaux simples, les entités nommées et certains types de pronoms ;
- les chunks prépositionnels (`XP`) : les `XN` précédés d'une préposition ;
- les chunks verbaux (cascade de 4 FSTs) : les voix actives et passives des infinitifs, participes passés, gérondifs et verbes conjugués (notés respectivement `XVI` – `XVIP` ; `XVK` – `XVKP` ; `XVG` – `XVGP` ; `XV` – `XVP`) ;

Les super-chunks héritent des propriétés morpho-syntaxiques de leur tête comme le montre la figure 3 qui représente un `XP`. `XP` hérite du lemme, du genre, du nombre et de la sous-catégorisation de sa tête (`^lemma`, `^gender`, `^number` et `^subcat`). Par ailleurs, nous conservons l'information liée à la préposition (`prep=$$.lemma`).

A la suite du processus de segmentation, l'automate du texte est nettoyé. La procédure de nettoyage consiste, d'une part, à supprimer les transitions dont les étiquettes n'appartiennent pas au niveau des super-chunks[7] (*e.g.* noms, verbes, adjectifs, ...) et, d'autre part, à conserver uniquement les chemins qui partent de l'état initial (début de phrase) et arrivent à l'état final (fin de phrase).

La procédure de segmentation en super-chunks appliquée à la séquence *au sujet d'un attentat terroriste* produit l'automate du texte donné à la figure 4.

[7] Notons toutefois que certaines entrées lexicales ne sont intégrées à aucun chunk (*i.e.* les conjonctions et les pronoms relatifs). Ces entrées sont conservées au même titre que les super-chunks.

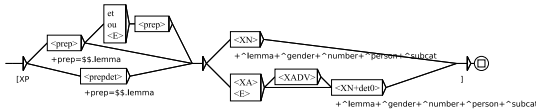


FIG. 3 – Chunk prépositionnel

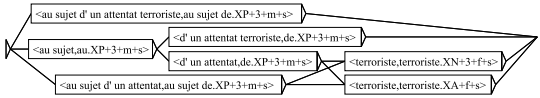


FIG. 4 – Automate du texte après segmentation

4 Levée d'ambiguïte incrémentale

La procédure de segmentation en super-chunks produit un ensemble d'analyses. Afin de réduire ou même de supprimer l'ambiguïté, le chunker inclut un module de levée d'ambiguïté composée de trois phases optionnelles : l'heuristique du plus court chemin, un jeu de règles et un module de décision statistique.

4.1 Application de l'heuristique du plus court chemin (SPH)

L'heuristique du plus court chemin consiste à ne garder, dans l'automate du texte, que les chemins les plus courts. Cette heuristique, indépendante de la langue, peut paraître simple et naïve au premier abord. Mais, en pratique, elle est très efficace. Elle privilégie en effet les analyses intégrant une ou plusieurs unités polylexicales au détriment des analyses compositionnelles. L'algorithme SPH est une adaptation de l'algorithme de Dijkstra (Dijkstra, 1959) qui garde l'ensemble des plus courts chemins d'un graphe au lieu d'un seul.

L'application de cette heuristique sur l'automate de la figure 4 du texte produit un automate totalement linéarisé : *<au sujet d'un attentat terroriste.XP>*.

4.2 Application de règles écrites manuellement

La plupart des ambiguïtés lexicales peuvent se résoudre efficacement en considérant leur contexte d'apparition. Dans cette perspective, nous avons développé un formalisme simple : les règles *Lubéron*. Une règle se compose de trois éléments : deux contextes (gauche et droit), éventuellement vides (`EMPTY`), représentés sous la forme de grammaires locales et une partie centrale listant une suite d'analyses ambiguës. Chaque règle décrit donc une ambiguïté potentielle[8]. Si cette dernière s'observe au sein de l'automate du texte, nous conservons uniquement la première analyse de la liste des éléments ambigus. Les autres analyses sont alors supprimées de l'automate.

`XN.wrtn`

[8]Le chunker contient actuellement 26 règles de ce type.

```
<XP> <XN>
EMPTY
```

La règle proposée ci-dessus exprime la contrainte suivante : dans le cas d'une ambiguïté XN – XP, l'analyse XP sera préférée si le contexte gauche (défini au sein du graphe XN.wrtn) présente un XN. Appliquée à l'automate de la figure 5, cette règle nous permet de supprimer l'analyse XN pour la séquence *de lutte contre le terrorisme*.

FIG. 5 – Ambiguïté des analyses en super-chunks

4.3 Application de règles statistiques simples

Certaines ambiguïtés ne peuvent être résolues efficacement par étude des contextes directs gauche ou droit. Un exemple prototypique est l'ambiguïté XV – XN (*e.g.* le mot *avions*, V (avoir) ou N (avion)). Dans ce cas, nous utilisons des règles de priorités statistiques apprises automatiquement au départ d'un corpus[9]. Étant donné un mot ambigu, l'analyse hors contexte la plus fréquente est choisie. Pour la forme *avions*, par exemple, nous retiendrons l'analyse N (probabilité de $0,6$) et supprimerons l'analyse V (probabilité de $0,4$). Si une forme ambiguë est absente de notre liste de décision, nous retenons, en dernier recours, la catégorie de (super-)chunk la plus fréquente.

Notons que toutes les phases de levée d'ambiguïté sont optionnelles. En effet, dans l'optique d'une analyse syntaxique, il peut être préférable de conserver une partie de cette ambiguïté, sa résolution pouvant entraîner des erreurs. L'ambiguïté XV – XN constitue, selon nous, une situation caractéristique qu'il est préférable de résoudre au niveau de l'attachement syntagmatique surtout si celui-ci est basé sur des règles lexicales.

5 Évaluation et discussion

La notion de super-chunk compatible avec notre définition n'existant dans aucun corpus annoté de référence, l'évaluation a dû être réalisée manuellement.

Notre procédure d'évaluation a porté sur un corpus composé de dépêches journalistiques extraites du site yahoo.fr. Ce corpus de 13 493 mots (*i.e.* 6 901 chunks), auxquels nous avons appliqué notre chunker à l'aide des données lexicales décrites dans les sections précédentes. La sortie est un texte annoté ne contenant plus aucune ambiguïté. Les résultats de l'évaluation sont donnés dans la table 2.

De manière générale, nous avons observé que la plupart des erreurs sont dues à l'incomplétude de nos ressources lexico-syntaxiques. Ceci implique que de nombreuses améliorations peuvent être apportées facilement et le seront très prochainement. Dans cette évaluation, nous distinguons les erreurs de précision et de rappel.

[9]Notre corpus comprend un an d'articles du journal Le Monde et a été étiqueté avec *TreeTagger* (Schmid, 1994).

PRÉCISION	RAPPEL	F MESURE
92,91 %	98,69 %	95.71%

TAB. 2 – Résultats

Les erreurs de **rappel** sont uniquement dues à la couverture de notre dictionnaire et de nos grammaires locales lexicalisées. D'un point de vue lexical, certains mots grammaticaux composés sont absents du dictionnaire (*e.g. tandis que*, *au-dessous de*). D'un point de vue plus grammaticale, les erreurs sont principalement imputables à la formalisation des entités nommées. La séquence *Nouri al Maliki*, par exemple, n'a pu être reconnue : la forme *al* est inconnue et n'a pas été intégré dans la grammaire comme un préfixe de nom de famille. Par ailleurs, quelques expressions semi-figées ont été oubliées (*e.g. vers 8h45*) de même que certaines structures pronominales complexes (*e.g. au cours de laquelle*).

Les erreurs de **précision** peuvent être divisées en quatre classes.

1. Erreurs liées à SPH

L'ambiguïté lexicale peut conduire a une mauvaise limitation des chunks après application de l'heuristique des plus courts chemins. Par exemple, dans la séquence *après l'affirmation du quotidien espagnol El Pais*, il existe deux analyses possibles :

– *[après l'affirmation XP] [du quotidien espagnol XP] [El Pais XN]* ;
– *[après l'affirmation XP] [du quotidien XP] [espagnol XA] [El Pais XN].*

Comme *quotidien* et *espagnol* peuvent être tous deux soit adjectif soit nom, l'algorithme SPH va préférer l'analyse `[Prep XA N]` au lieu de `[Prep N] [XA]`.

2. Erreurs dues aux règles statistiques

Dans la séquence *La côte Est et les villes de New York ...*, deux analyses sont attribuées au chunk *Est* : il s'agit soit d'un `XV` (être), soit d'un `XA` (direction est). Bien qu'*Est* soit `XA` dans ce contexte, le module statistique va préférer l'analyse `XV` (probabilité de $0,9$ contre $0,1$ pour l'analyse *XA*).

3. Erreurs causées par l'application des règles Luberon

Ces erreurs sont heureusement très rares. Elles concernent principalement l'ambiguïté `XP–XN` due à la forme *de* qui peut être déterminant et préposition. Dans la séquence *qui n'a pas fourni de plus amples détails*, par exemple, le chunk *de plus amples détails* aurait dû être étiqueté *XN*.

4. Erreurs imputables à la couverture lexicale

Quelques structures composées absentes dans le dictionnaire provoque des erreurs. Par exemple, *en outre* est un adverbe composé mais est absent de notre dictionnaire. Ainsi, l'analyse compositionnelle est choisie dans la phrase *ils ont en outre pris plusieurs centaines de personnes en otage*. Elle est segmentée en super-chunks de la manière suivante :

– *[ils XN] [ont XV] [en outre pris plusieurs centaines XP] [de personnes XP] [en otage XP]*

au lieu de,

– *[ils XN] [ont en outre pris XV] [plusieurs centaines XN] [de personnes XP] [en otage XP]*

où *en outre* est un adverbe inséré dans un chunk verbal.

En plus de l'évaluation en rappel et précision, nous avons aussi estimé l'impact des unités polylexicales pour l'attachement lexical. Notre procédure permet la réalisation correcte de 36,6 % des attachements lexicaux intérieurs aux groupes nominaux et prépositionnels, soit environ 13 % des attachements internes et externes aux syntagmes.

Malgré ces quelques erreurs, notre évaluation montre, selon nous, l'intérêt d'une segmentation en super-chunks, tant du point de vue de l'attachement que du point de vue de la réduction globale de l'ambiguïté.

6 Conclusion et perspectives

Dans cette article, nous avons présenté une technique de *chunking* reposant sur une augmentation significative du niveau lexical. En introduisant la notion de *super-chunks*, nous cherchions, d'une part, à optimiser le processus de désambiguïsation et, d'autre part, à résoudre une part de l'attachement lexical au sein des constituants prépositionnels et nominaux.

Afin d'évaluer la pertinence et l'efficacité de notre hypothèse, nous avons confronté notre chunker à un corpus de dépêches journalistiques. Cette expérience nous a permis de dégager une double conclusion.

– Notre procédure affiche une précision et un rappel excellents sans nécessiter le recours à un étiqueteur.
– La prise en compte des unités polylexicales nous permet d'évacuer efficacement (*i.e.* sans entraîner d'erreurs) une part conséquente des attachements internes aux constituants nominaux et prépositionnels.

Cette expérience nous a également permis de préciser un certain nombre de perspectives organisant notre travail futur. Ces perspectives s'articulent autour de deux points principaux : (1) l'augmentation des ressources lexicales (principal facteur de succès de notre application) et (2) l'amélioration du module de désambiguïsation statistique (en ce sens, l'intégration des HMM nous semble être une solution intéressante).

Références

ABNEY S. P. (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, **2**(4), 337–344.

AIT-MOKHTAR S. & CHANOD J.-P. (1997). Incremental finite-state parsing. In *Proceedings of the fifth Conference on Applied Natural Language Processing ANLP'97*.

BLANC O. & CONSTANT M. (2006). Outilex, a linguistic platform for text processing. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, p. 73–76.

COURTOIS B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, **87**, 11–22.

COURTOIS B., GARRIGUES M., GROSS G., GROSS M., JUNG R., MATHIEU-COLAS M., MONCEAUX A., PONCET-MONTANGE A., SILBERZTEIN M. & VIVÈS R. (1997). *Dictionnaire électronique DELAC : les mots composés binaires*. Rapport interne, LADL (Paris 7).

DAILLE B. (1995). *Combined approach for terminology extraction : lexical statistics and linguistic filtering*. Rapport interne, Lancaster University.

DIJKSTRA E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

FEDERICI S., MONTEMAGNI S. & PIRELLI V. (1996). Shallow parsing and text chunking : A view on underspecification in syntax. In *Proceedings of the ESSLLI'96 Workshop on Robust Parsing*.

GROSS M. (1997). *The construction of local grammars*, p. 329–352. MIT Press : Cambridge.

JOSHI A. & HOPELY P. (1997). A parser from antiquity : an early application of finite state transducers to natural language parsing. *Natural Language Engineering*, **2**(4), 6–15.

KARLSSON F., VOUTILAINEN A., HEIKKILA J. & ANTTILA A. (1995). *Constraint Grammar : A language-independent system for parsing unrestricted text*, volume 4 of *Natural Language Processing*. Mouton de Gruyter.

NIVRE J. & NILSSON J. (2004). Multiword units in syntactic parsing. In *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*, p. 39–46, Lisbon.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

SERETAN V., NERIMA L. & WEHRLI E. (2003). Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of th* 4^{th} *International Conference on Recent Advances in NLP (RANLP-2003)*, p. 424–431.

WATRIN P. (2006). *Une approche hybride de l'extraction d'information : sous-langages et lexique-grammaire*. PhD thesis, Université catholique de Louvain.

WOODS W. A. (1970). Transition network grammars for natural language analysis. *Communications of the ACM*, **13**(10).

Détection et prédiction de la satisfaction des usagers dans les dialogues Personne-Machine

Narjès BOUFADEN, Truong LE HOANG, Pierre DUMOUCHEL
Centre de Recherche Informatique de Montréal
et École de technologie supérieure
{Narjes.Boufaden,LeHoang.Truong,Pierre.Dumouchel}@crim.ca

Résumé. Nous étudions le rôle des entités nommées et marques discursives de rétroaction pour la tâche de classification et prédiction de la satisfaction usager à partir de dialogues. Les expériences menées sur 1027 dialogues Personne-Machine dans le domaine des agences de voyage montrent que les entités nommées et les marques discursives n'améliorent pas de manière significative le taux de classification des dialogues. Par contre, elles permettent une meilleure prédiction de la satisfaction usager à partir des premiers tours de parole usager.

Abstract. We study the usefulness of named entities and acknowldgment words for user satisfaction classification and prediction from Human-Computer dialogs. We show that named entities and acknowledgment words do not enhance baseline classification performance. However, they allow a better prediction of user satisfaction in the beginning of the dialogue.

Mots-clés : prédiction de la satisfaction usager, classification des dialogues Personne-Machine.

Keywords: prediction of user satisfaction, Human-Computer dialog classification.

1 Introduction

La progression des systèmes de dialogue Personne-Machine dans le marché du service à la clientèle crée des attentes grandissantes tant sur le plan de la gestion des données générées par ces systèmes que sur leur exploitation à des fins d'évaluation.

La classification, l'indexation et l'extraction d'information sont autant d'exemples d'applications peu ou pas encore explorées en gestion des dialogues Personne-Machine. La majeure partie de la recherche dans ce domaine est encore consacrée à l'évaluation de ces systèmes.

Dans cet article, nous explorons la classification des dialogues Personne-Machine dans le but de détecter les dialogues problématiques dans lesquels l'usager montre une insatisfaction par rapport au système. Nous explorons l'utilisation des entités nommés et des marques discursives de rétroaction pour la tâche de prédiction de la satisfaction usager durant et après le déroulement du dialogue.

Ces travaux s'insèrent dans le cadre d'un projet en cours avec une compagnie de télécommu-

nication dans le but d'évaluer leur système de dialogue et analyser les dialogues qu'il génère. Comme première étape de ce projet, nous étudions la problématique de détection et prédiction de la satisfaction usager sur un corpus public : le DARPA Communicator.

Dans la section 2, nous présentons l'état de l'art en évaluation des systèmes de dialogues Personne-Machine, en détection des dialogues problématiques et présentons le cadre théorique PARADISE. Dans la section 3, nous présentons le corpus DARPA Communicator avec lequel nous effectuons nos expériences. La section 4 décrit notre approche basée du cadre théorique PARADISE. La section 5 présente trois expériences de classification dans lesquelles nous expérimentons différentes combinaisons des attributs proposés pour détecter la satisfaction usager à partir du dialogue. Dans la section 6, nous étudions le rôle des entités nommées et des marques discursives dans la prédiction de la satisfaction usager en début de dialogue.

Enfin, la section 7 présente une synthèse de nos résultats ainsi que les prochaines étapes de ce projet.

2 État de l'art

L'évaluation des systèmes de dialogue occupe depuis la fin des années 90 une place de plus en plus importante en recherche. Par exemple, (Eckert *et al.*, 1998) ont utilisé un modèle stochastique pour modéliser le comportement de différentes classes d'usager dans le but de collecter des statistiques sur les différents scénarii de dialogues. En particulier, ils ont montré que la longueur du dialogue en termes du nombre de tour de parole est un bon prédicteur des performances d'un système.

Dans une perspective plus générale, (Walker *et al.*, 1997) proposaient le cadre théorique PARADISE actuellement le plus utilisé pour l'évaluation des systèmes de dialogues. Ce cadre théorique s'inspire de la théorie de décision et pose comme hypothèse que la performance d'un système de dialogues est corrélée avec le degré de convivialité. Dans le cadre de notre application (le service à la clientèle), la convivialité se mesure en termes de satisfaction de l'usager.

PARADISE met en avant la maximisation de la satisfaction usager en maximisant les chances de réussir la tâche ou but du dialogue tout en minimisant les coûts associés à sa réalisation. Ces coûts sont définis en termes d'efficacité (i.e. nombre de tours de parole système et usager) et qualité du système (i.e. temps de réponse du système). La figure 1 illustre le cadre théorique PARADISE.

Le cadre PARADISE a été utilisé dans plusieurs travaux (Walker *et al.*, 2000),(Lamel & Rosset, 2000),(Devillers & Rosset, 2000), notamment dans le but de choisir les stratégies de dialogues qui maximisent la satisfaction de l'usager (Walker & Passonneau, 2001), pour déterminer les composantes (reconnaissance de la parole, stratégie de dialogue) ayant le plus d'impact sur la performance d'un système ou encore détecter les dialogues problématiques (Hastie *et al.*, 2002).

En particulier, (Hastie *et al.*, 2002) ont proposé une approche pour la détection des dialogues problématiques utilisant 16 attributs représentant différentes mesures reliées aux trois dimensions définies dans PARADISE, à savoir :

Mesure du succès de la tâche Évaluation de l'usager indiquant la réalisation de la tâche.

Mesure de l'efficacité du système Traits extraits des logs du système de dialogue.

- Manuellement : Taux d'erreur de reconnaissance de la parole calculé à partir des trans-

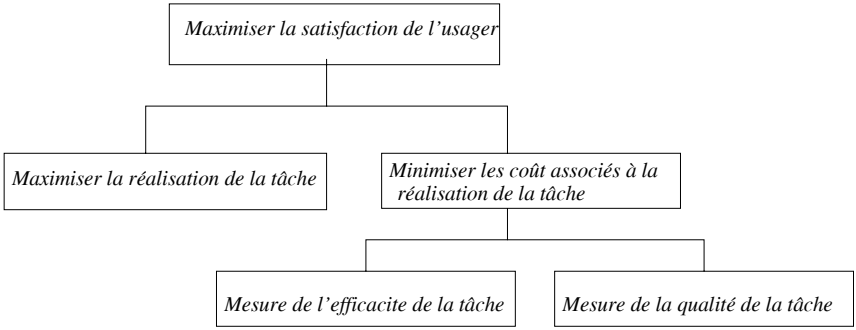


FIG. 1 – Cadre théorique PARADISE

criptions manuelles des dialogues et le taux d'erreur phrastique.
– Automatiquement : la durée de la tâche, le nombre de tours de parole durant la tâche, nombre de chevauchement de tours de parole système et usager, la moyenne des temps des tours de parole usager, la moyenne des mots par tour de parole usager, la moyenne des temps des tours de parole système, la moyenne des mots par tour de parole système et le type de téléphone (cellulaire ou fixe).

Mesure de la qualité du système Actes de dialogue associés aux tours de parole du système.

Dans une première expérience, les auteurs ont utilisé tous les traits : extraits automatiquement ainsi que ceux annotés manuellement. En utilisant un arbre de décision ils ont obtenus un taux de classification de 54%. Toutefois, pour éviter l'utilisation des attributs annotés manuellement, ils ont entraîné un arbre de décision afin de prédire le **succès de la tâche** et les actes de dialogues des tours de parole système. Avec un taux de prédiction du **succès de la tâche** de 92% et un taux de prédiction des actes de dialogues de 98%, les auteurs ont obtenus un taux de classification similaire.

(Walker *et al.*, 2002) ont testé la détection des dialogues problématiques en utilisant des attributs entièrement extraits de manière automatique. Ils ont utilisé aussi des mesures de l'efficacité modélisé avec l'algorithme RIPPER (Cohen, 1995) : un algorithme de classification à base de règles. Sur le corpus généré par le système de AT&T *How May I Help You*, les auteurs ont obtenu un taux de classification de 70,1% à partir du premier tour de parole usager, 78,4% à partir des deux premiers tours de parole et de 83% sur tout le dialogue.

Les travaux que nous présentons se basent sur le cadre théorique PARADISE et s'inspirent des travaux de (Hastie *et al.*, 2002) réalisés sur le corpus DARPA Communicator.

3 Corpus DARPA Communicator

Le corpus DAPRA Communicator (version 2001) est un corpus public distribué par le Linguistic Data Consortium (LDC). C'est le résultat d'une expérience menée sur plusieurs sites incluant 8 systèmes de dialogue Personne-Machine dans le but de développer des approches robustes de reconnaissance de la parole et de gestion de dialogues pour l'accès interactif à l'information (Robust Recognition and Dialog Tracking for Interactive Information Access). Le

domaine d'application choisi est celui des agences de voyages.

Un des défis de cette expérience était de collecter un corpus de dialogues Personne-Machine proche de la réalité. Pour ce faire, les scénarii des dialogues étaient en majorité prédéfinis (sept des neuf scénarii) puisque chaque participant savait l'origine et la destination de son voyage ainsi que les dates et la compagnie aérienne à choisir. Tandis que deux des scénarii étaient laissés au choix des participants. La version 2001 du corpus contient une sélection de plus de 1242 dialogues annotés avec la degré de satisfaction de l'usager. Les dialogues sont en moyenne composés de 51 tours de parole avec en moyenne 25,4 tours de parole usager et une longueur moyenne d'un tour de parole usager est de 64,6 mots.

Dans la version 2001 du corpus que nous utilisons dans nos expériences, chaque dialogue est accompagné de deux fichiers :
– Un fichier qui contient la transcription du dialogue (Figure 2 partie du haut).
– Un fichier récapitulatif contenant les scores attribués par l'usager pour l'évaluation de la convivialité du système (Figure 2 partie du bas). En allant de gauche à droite la ligne contient le nom du système (CMU pour Carnegie Melon University), l'indexe du site (le chiffre 27), le temps de début du dialogue (14 :25), la date du dialogue (2000/07/06), la complétion de la tâche et le reste des données sont respectivement les réponses à des questions portant sur la convivialité du système. Chaque question était notée sur une échelle de 1 à 5 avec la valeur 1 indiquant que l'usager est en parfait accord avec l'affirmation et le score 5 indiquant un total désaccord.

Thu Jul 6 2000 at 15 :15 :38.51 : Task-specific portion started. Thu Jul 6 2000 at 15 :17 :01.44 : Overall task started. Task completion status : not completed. Thu Jul 6 2000 at 15 :15 :24.85 to Thu Jul 6 2000 at 15 :15 :25.01 : New system turn began. Thu Jul 6 2000 at 15 :15 :24.94 : System started speaking. Thu Jul 6 2000 at 15 :15 :36.22 : System finished speaking. System said : . Hello. Welcome to the C M U Communicator. Please speak your 4-digit ID number using the phrase, . My ID number is Thu Jul 6 2000 at 15 :15 :38.51 : New user turn began. Thu Jul 6 2000 at 15 :15 :38.51 : User started speaking. Thu Jul 6 2000 at 15 :15 :43.53 : User finished speaking. Recognizer heard : MY I D NUMBER IS . ?THAT ?. . ?WONDER ?. ZERO EIGHT SEVEN User said : my i. d. number is one zero eight seven [h#] Figure 2 : Exemple de fichier résumé associé au dialogue 1087_02_27_03_20000706
CMU 27 14 :25 EDT 2000/07/06 Alive No 4 1 4 3 4 (no comments provided)

FIG. 2 – Exemple des logs du système extrait du corpus DARPA Communicator

4 Approche

Nous proposons d'utiliser PARADISE pour détecter les dialogues problématiques et prédire la satisfaction usager durant et après le déroulement du dialogue.

Dans notre approche nous tenons compte de deux contraintes liées à notre projet :
– Minimiser la quantité d'information à annoter manuellement. Le but de ces expériences étant de concevoir une application réelle pour la détection des dialogues problématiques, nous voulons obtenir un système entièrement automatisé.
– Minimiser la quantité d'information dépendante du domaine. Notre système étant voué à être appliqué sur un corpus différent du DARPA Communicator, nous voulons minimiser l'information dépendante du domaine des agences de voyages.

Nous proposons de partir des **mesures d'efficacité** proposées par (Hastie *et al.*, 2002). Plus précisemment, nous partons des mesures extraites automatiquement et étudions deux nouveaux traits indépendants du domaine : les **entités nommées** contenue dans un dialogue et les **marques discursives de rétroaction**. Nous proposons d'utiliser les **entités nommées** comme indicateur de la densité d'information que nous corrélons avec la qualité du dialogue. Nous pensons qu'une densité d'information faible (interruption du dialogue) ou trop importante (plusieurs répétitions) pourrait indiquer un problème de communication entre l'usager et le système. Les **marques discursives de rétroaction** sont utilisées dans la détection des actes de dialogues notamment d'acquiescement ou de confirmation, tous les deux correlés avec la satisfaction de l'usager (Colineau & Caelen, 1996).

4.1 Les entités nommées

Les entités nommés sont des noms propres ou chiffres référant à des noms de personnes, de lieux ou des noms de compagnies. L'apport des entités nommées pour les applications de compréhension du langage naturel a été largement démontré aussi bien en extraction d'information qu'en résumé automatique ou en traduction automatique. Leur intérêt réside dans la valeur sémantique de l'information qu'ils véhiculent. Dans le cadre de notre application, nous nous intéressons à la correlation existant entre les entités nommées et la densité d'information.

Nous proposons d'utiliser le compte des entités nommées comme indicateur de la densité d'information contenue dans un dialogue. Par ailleurs et dans la mesure où notre corpus n'est pas de taille importante, nous regroupons le compte de toutes les catégories d'entités nommées pour réduire la dimension des vecteurs des données fournies comme donnée d'entraînement.

4.2 les marques discursives de rétroaction

Les marques discursives de rétroaction telles que `ok`, `yes` et `no` sont des marques lexicales utilisées pour détecter les actes de dialogues d'acquiescement ou d'opposition (Colineau & Caelen, 1996), (Jurafsky *et al.*, 1998). Ce sont des unités lexicales qui dans le contexte des dialogues Personne-Machine donnent une mesure qualitative sur le déroulement du dialogue.

Par ailleurs, les stratégies dialogiques des systèmes de dialogues Personne-Machine étant en grande partie composé de question à base de choix binaires (Yes/No questions) ou de choix multiples, nous sommes assurés d'observer ces marques de manière significative dans les dialogues. Dans notre approche, nous ne faisons aucune distinction entre les marques de rétroactions indiquant un acquiescement `yes`, `ok`, `correct` ou `yep` de celles indiquant une opposition `no`, `wrong` ou `erase` puisque nous ne tenons pas compte du contexte précédent le pour de parole. Aussi, ce choix nous permet aussi de réduire la dimension des vecteurs des données d'entraînement.

Dans ce qui suit, nous testons différentes combinaisons des attributs décrits dans notre approche pour la tâche de classification des dialogues et la prédiction de la satisfaction usager durant le déroulement du dialogue.

5 Expérience 1 : Détection des dialogues problématiques

Notre première expérience a pour but de classer des dialogues dans une des classes suivantes :
– Positive : L'usager a attribué un score de satisfaction < 12. Ce score est obtenu en cumulant tous les scores des cinq questions évaluant la convivialité du système. Le seuil de 12 est aligné sur celui proposé dans les travaux de (Hastie *et al.*, 2002) afin de permettre la comparaison de nos résultats.
– Négative : L'usager a attribué un score ≥ 12.

Afin d'évaluer notre approche, nous comparons nos résultats avec ceux de (Hastie *et al.*, 2002) en utilisant que les mesures d'efficacité (indépendantes du domaine) extraites de manière automatique et les actes de dialogues (dépendants du domaine) pour classer les dialogues.

Nous présentons quatre expériences combinant nos différents attributs :

Eff. composé uniquement des mesures de l'efficacité extraites automatiquent à partir des logs du système de dialogue.

Eff.+NE composé des mesures de l'efficacité et du compte des entités nommées toutes catégories confondues.

Eff.+ACK composé des mesures de l'efficacité et du compte des marques discursives de rétroaction.

Eff.+EN+ACK composé de tous les attributs.

Nous testons ces combinaisons avec trois algorithmes de classification : Support Vector Machine (SVM), k-Nearest-Neighbour (kNN) et un arbre de décision (DT). Le corpus utilisé est constitué de 1027 dialogues tirés du corpus DARPA Communicator 2001. À la base le corpus contenait une distribution de 4 dialogues de la classe positive pour un dialogue de la classe négative. Les premiers résultats sur ce corpus avec une distribution très biaisée en faveur des dialogues positifs se rapprochaient sensiblement du baseline de 80%. Afin de remédier à ce biais, nous avons constitué un nouveau corpus à partir du corpus original en dupliquant de manière aléatoire les statistiques de dialogues de la classe négative et en retirant de manière aléatoire les statistiques de dialogues de la classe positive jusqu'à obtention d'une distribution uniforme (resampling avec duplication) avec 50% des données par classes.

Les résultats que nous présentons sont obtenus sur ce nouveau corpus. Ce sont les moyennes des résultats de 10 validations croisées.

Modele	Baseline	Eff.	Eff.+NE	Eff.+ACK	Eff.+NE+ACK	(Hastie *et al.*, 2002)
SVM	50%	61,26%	62,9%	62,83%	63,51%	
kNN	50%	91,41%	91,12%	91,6%	91,8%	
DT	50%	85,75%	84,68%	87,41%	87,26%	54%

TAB. 1 – Performance des différents classificateurs en termes de taux de classification pour les différentes combinaisons d'attributs.

Contrairement à notre intuition quant à l'apport des entités nommées et des marques discursives, aucune amélioration significative est observée pour les différents modèles. Les résultats pour le

modèles kNN est sensiblement le même pour toutes les combinaisons d'attributs. Tandis qu'une petite amélioration est observée pour l'arbre de décision (DT) et le SVM. Cependant, le résultat obtenu pour le kNN avec la combinaison de tous les attributs est supérieur aux résultats obtenu par (Hastie *et al.*, 2002). Toutefois, rappelons que la distribution des classes n'étant pas la même ont ne peut établir une comparaison directe entre nos résultats et les leurs.

Enfin, bien que nous n'ayons pas amélioré le résultat de la classification obtenu avec les mesures d'efficacité, nous avons obtenu un meilleur taux de classification de 91,8% avec le modèle kNN. Pour ce modèle les pourcentages de précision et de rappel par classe sont illustrés dans le Tableau 2.

Classe	Précision	Rappel	F-score
Positive	84,4%	90,1%	89,6%
Négative	90,5%	84,9%	89,6%

TAB. 2 – Rappel, précision et moyenne harmonique (F-score) des classes Positive et Négative pour le modèle kNN.

6 Expérience 2 : Prédiction de la satisfaction de l'usager

Malgré le peu d'amélioration des résultats obtenus sur la classification des dialogues en combinant toutes les marques, nous voulions tester l'apport des entités nommés et des marques discursives pour la prédiction de la satisfaction usager durant le déroulement du dialogue. Nous avons conduit neuf expériences dans lesquelles nous testons successivement les différentes combinaisons des attributs du tableau 1 avec des données issues uniquement d'un nombre variable de tours de parole usager. Chacune des combinaisons est testée sur des parties d'un dialogue comprenant respectivement 1 tour de parole de l'usager, 2 tours de parole et ce jusqu'à 8 tours de parole et enfin sur tout le dialogue.

En augmentant le nombre des tours de parole de l'usager à chaque nouvelle expérience, nous avons dressé des courbes illustrant la progression du taux de classification en fonction de la progression du dialogue. Nous avons utilisé le même corpus que celui utilisé dans la première expérience et les résultats obtenus sont la moyenne de 10 validations croisées.

Les courbes obtenues sont montrées dans les figures 3, 4 et 5. La première figure représente l'évolution du taux de classification en fonction du nombre de tours de parole usager et ce pour différentes combinaisons des attributs modéliser par un SVM. La deuxième figure montre la progression modélisée avec l'algorithme kNN et la dernière figure montre la progression modélisée avec un arbre de décision.

Nous remarquons sur les différents graphiques que pendant les trois premiers tours de parole usager, la combinaison de tous les attributs donne une meilleure prédiction de la satisfaction usager.

Aussi, nous remarquons que le meilleur résultat est en majorité obtenu pour la combinaison Eff.+ACK. Les marques discursives combinées aux mesures d'efficacité améliorent grandement le taux de classification pendant les premiers tours de parole et ce pour tous les modèles. Il semble que l'ajout des entités nommées introduisent du bruits car le résultat obtenu pour la combinaison Eff.+EN+ACK est moins bon.

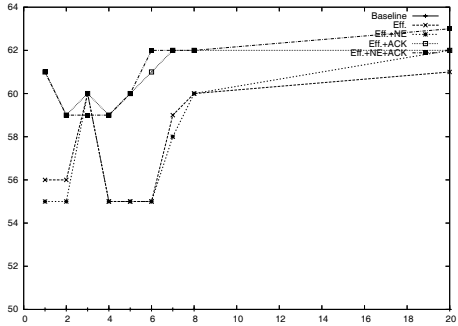


FIG. 3 – Courbes illustrant la progression du taux de classification durant le dialogue pour le modèle SVM.

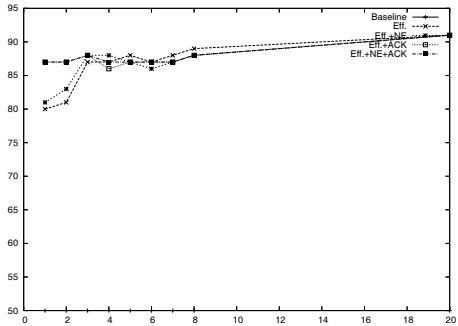


FIG. 4 – Courbes illustrant la progression du taux de classification durant le dialogue pour le modèle kNN.

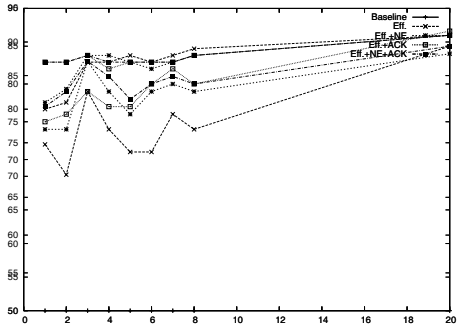


FIG. 5 – Courbes illustrant la progression du taux de classification durant le dialogue pour le modèle DT.

Par ailleurs, la meilleure performance est obtenue avec l'algorithme kNN qui se base sur la distance euclidienne pour classer les dialogues. Cela s'explique par le fait que les données d'entraînement sont des valeurs réelles qui représentent des fréquences et que la distance euclidienne est une fonction qui permet de représenter efficacement la similarité en termes de proximité.

Le tableau 3 montrent les performances du modèle kNN qui a donné le meilleur résultat de classification avec les différentes combinaisons d'attributs.

Nb tours de parole	Eff.	Eff.+EN	Eff.+ACK	Eff.+EN+ACK
1	80,49%	81,78%	87,07%	87,26%
2	81,47%	83,44%	87,65%	87,26%
3	87,80%	88,29%	88,48%	88,39%
tous	91,41%	91,12%	91,6%	91,8%

TAB. 3 – Progression du taux de classification sur les trois premiers tours de parole usager pour le modèle kNN.

7 Conclusion

Nous avons testé la combinaison des mesures de l'efficacité proposées par (Hastie *et al.*, 2002) avec deux nouveaux attributs : les entités nommés et les marques discursives de rétroaction pour la classification des dialogues et la prédiction de la satisfaction usager. Bien que ces deux attributs n'aient pas améliorés de manière significative les résultats de la classification des dialogues, ils ont permit une meilleur prédiction de la satisfaction usager durant les premiers tours de parole de l'usager.

En particulier, le compte des marques discursives a permit d'obtenir les meilleurs taux de prédiction en début de dialogue. Ce résultat a un intérêt particulier puisque moyennant cet attribut facilement calculable et non dépendant du domaine d'application, nous avons amélioré le taux de prédiction de la satisfaction usager de 6,6% pour le premier et second tour de parole par rapport à celui obtenu avec le modèle utilisant un arbre de décision.

Dans les prochaines étapes, nous exploiterons d'avantage les marques discursives de rétroaction en tenant compte de l'information prosodique extraites de l'audio de ces unités lexicales. L'ajout de la prosodie permettra de distinguer ces marques en leur associant un contenu émotionnel pour une meilleur prédiction de la satisfaction usager.

Remerciement

Ce projet est rendu possible grâce au support de Patrimoine Canada.

Références

COHEN W. (1995). Fast Effective rule induction. In *Proceedings of the 12th Conference on Machine Learning*.

COLINEAU N. & CAELEN J. (1996). Une approche lexicale pour la reconnaissance d'actes de dialogue. In *Séminaire lexique en traitement automatique de la parole*, p. 137–145, Toulouse, France.

DEVILLERS B.-M. H. L. & ROSSET S. (2000). Predictive performance of dialog systems. In *Int. Conf. on Language ResourcesandEvaluation,LREC2000*.

ECKERT W., LEVIN E. & PIERACCINI R. (1998). Automatic evaluation of spoken dialogue systems. *TWLT13 : Formal semantics and pragmatics of dialogue*.

HASTIE H., PRASAD R. & WALKER M. (2002). What's the Trouble : Automatically Identifying Problematic Dialogues in DARPA Communicator Dialogue Systems. In *Proceedings of the ACL 2002*.

JURAFSKY D., E. S., B. F. & CURL T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING 98 Workshop on Discourse Relations and Discourse Markers*.

LAMEL L. & ROSSET S. (2000). Considerations in the design and evaluation of spoken language dialog systems. In *ICSLP, 2000*.

WALKER M., LITMAN D. J., KAMM C. A. & ABELLA A. (1997). Evaluating interactive dialogue systems : Extending component evaluation to integrated system evaluation. In J. HIRSCHBERG, C. KAMM & M. WALKER, Eds., *Interactive Spoken Dialog Systems : Bridging Speech and NLP Together in Real Applications*, New Brunswick, New Jersey : ACL.

WALKER M. A., LANGKILDE I., WRIGHT J., GORIN A. & LITMAN D. (2000). Learning to predict problematic situations in a spoken dialogue system : Experiments with how may i help you ? In *North American Meeting of the Association of Computational Linguistics*.

WALKER M. A., LANGKILDE-GEARY I., HASTIE H. W., WRIGHT J. & GORIN A. (2002). Automatically Training A Problematic Dialog Predictor for the HMOHY Spoken Dialog System. *Journal of Artificial Intelligence Research*, **16**, 293–319.

WALKER M. A. & PASSONNEAU R. (2001). Date : A dialogue act tagging scheme for evaluation of spoken dialogue systems. In *In Proceedings of Human Language Technology Conference*, San Diego.

Les ellipses dans un système de traduction automatique de la parole

Pierrette BOUILLON[1], Manny RAYNER[1, 2]
Marianne STARLANDER[1], Marianne SANTAHOLMA[1]
[1] University of Geneva, TIM/ISSCO
40, bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
[2] Powerset Inc, 475 Brannan Street, San Francisco, CA 94107, US
{Pierrette.Bouillon,Emmanuel.Rayner}@issco.unige.ch
{Marianne.Starlander,Marianne.Santaholma}@eti.unige.ch

Résumé. Dans tout dialogue, les phrases elliptiques sont très nombreuses. Dans cet article, nous évaluons leur impact sur la reconnaissance et la traduction dans le système de traduction automatique de la parole MedSLT. La résolution des ellipses y est effectuée par une méthode robuste et portable, empruntée aux systèmes de dialogue homme-machine. Cette dernière exploite une représentation sémantique plate et combine des techniques linguistiques (pour construire la représentation) et basées sur les exemples (pour apprendre sur la base d'un corpus ce qu'est une ellipse bien formée dans un sous-domaine donné et comment la résoudre).

Abstract. Elliptical phrases are frequent in all genres of dialogue. In this paper, we describe an evaluation of the speech understanding component of the MedSLT medical speech translation system, which focusses on the contrast between system performance on elliptical phrases and full utterances. Ellipsis resolution in the system is handled by a robust and portable method, adapted from similar methods commonly used in spoken dialogue systems, which exploits the flat representation structures used. The resolution module combines linguistic methods, used to construct the representations, with an example-based approach to defining the space of well-formed ellipsis resolutions in a subdomain.

Mots-clés : traduction automatique de la parole, reconnaissance de la parole, ellipses, évaluation, traitement du dialogue, modèle du language fondé sur les grammaire.

Keywords: speech recognition, speech translation, ellipsis, dialogue processing, grammar-based language modelling, evaluation.

1 Introduction

Dans tout dialogue, les phrases elliptiques sont très nombreuses et il semble important de pouvoir les traiter correctement (Fernandez & Ginzburg, 2002). Ceci est particulièrement vrai avec MedSLT, un système de traduction automatique de la parole dans le domaine médical. Celui-ci traduit des questions de diagnostic pour des patients étrangers, et ceci en anglais, français, japonais, espagnol, catalan et arabe (Bouillon *et al.*, 2005). Dans sa version unidirectionnelle (où le médecin pose essentiellement des questions de type oui-non auxquelles le patient répond

de manière non verbale), l'utilisation des ellipses permet en effet au docteur (M) d'éviter les répétitions et pallie ainsi la nécessité de poser des question oui-non (1)[1] :

(1) M : avez-vous mal sur le côté droit ?
M : le côté gauche ?
M : des deux côtés ?
M : avez-vous ces douleurs depuis une semaine ?
M : deux semaines ?
M : plus de deux semaines ?

Dans la version bidirectionnelle (où le patient (P) répond aux questions du docteur), elle permet de simplifier la tâche. Si le système n'accepte que des réponses elliptiques, comme en (2), il devient en effet plus facile de prévoir les réponses du patient (il y moins de variations stylistiques possibles avec des phrases courtes que complètes) et de le guider ensuite vers des phrases couvertes par le système.

(2) M : do you have a fever ?
P : un poco / mucho / sí treinta y nueve

L'utilisation des ellipses dans ce contexte particulier soulève cependant deux questions : (1) qu'en est-il de la reconnaissance des ellipses, par rapport aux phrases complètes ? et (2) est-il possible, avec les techniques actuelles, d'obtenir une traduction de qualité suffisante pour ce type d'application ? La première question n'a pas encore reçu beaucoup d'attention à notre connaissance. A priori, les ellipses devraient avoir des conséquences bénéfiques pour la reconnaissance. Du fait que les phrases sont plus courtes, celle-ci devrait prendre moins de temps ; elle devrait aussi être de meilleure qualité - on s'attend à avoir moins d'erreurs dans une phrase courte que longue. Mais on sait par ailleurs que les phrases et mots courts peuvent être difficiles à reconnaître, faute de contraintes syntaxiques et sémantiques venant du contexte. Il semble donc important de vérifier leur impact réel sur la reconnaissance. Pour la traduction, différentes approches sont envisageables. Dans un système général, il devrait être possible de traduire les ellipses mot à mot. Mais si la qualité est importante, comme dans le domaine médical, les problèmes seront très nombreux et on peut se demander jusqu'à quel point ce type de traduction restera compréhensible (Boitet *et al.*, 2002). Même pour des langues proches et dans un domaine limité comme le nôtre, la traduction des ellipses ne peut pas se faire sans contexte. Dans l'exemple (3), celui-ci conditionne l'accord de l'adjectif.

(3) M : is the pain severe ?
Trad : la douleur est-elle intense ?
M : moderate
Trad : modéré / modérée / modérées / modérés

Il est aussi indispensable pour que les règles de traduction puissent s'appliquer correctement. En (4), *for* doit être traduit par "depuis" si l'aspect de la phrase est perfectif.

(4) M : have you had headaches for days ?
Trad : avez vous mal depuis plusieurs jours ?
M : for weeks
Trad : pendant plusieurs semaines

En (5), le complément de lieu en anglais (*in your stomach*) devient le sujet de la phrase en espagnol, ce qui fait que la traduction de l'ellipse devrait changer de catégorie syntaxique ("in

[1]Tous les exemples de cet article sont directement repris des données collectées dans le cadre du projet MedSLT.

your head" (pp) → *la cabeza* (np)). Ce type de divergence est fréquent aussi quand on traduit vers des langues plus éloignées. En japonais par exemple, les ellipses adjectivales ou nominales relèvent de la langue familière et se traduisent plutôt par des phrases complètes comme en (6).

(5) M : do you have a pain in your stomach ?
Trad : le duele el estomago ?
M : in your head
Trad : en la cabeza ?

(6) M : is the pain severe ?
Trad : hageshii itami desu ka ? (sevère douleur est Q ?)
M : moderate ?
Trad : *chuuteido / *chuuteido no (moderée / moderée GEN)
chuuteido no itami desu ka (modérée GEN douleur est Q)

Pour éviter la traduction mot à mot, une autre solution, à l'extrême opposé, consiste à effectuer une analyse profonde au niveau de la syntaxe et du discours, de manière à pouvoir traduire les ellipses en contexte. C'est l'optique choisie dans Verbmobil (Kipp *et al.*, 2000). On sait cependant que les différentes composantes sous-jacentes (mémoire du dialogue, processeur de plan et processeur de dialogue) sont très coûteuses et restent peu robustes et portables. Nous proposons donc plutôt une approche intermédiaire, possible elle aussi pour des domaines limités. Empruntée directement aux méthodes similaires traditionnellement utilisées en traitement de dialogue (par exemple, (Ward & Issar, 1994) dans le contexte de ATIS), elle exploite des représentations sémantiques plates (Rayner *et al.*, 2005) et combine des techniques linguistiques (pour construire la représentation) et d'apprentissage (pour apprendre ce qu'est une ellipse bien formée dans un sous-domaine donné et comment la résoudre). Dans la suite, nous décrivons d'abord MedSLT, en focalisant sur les composantes qui nous intéressent ici, puis nous présentons la manière dont les ellipses y sont traitées. L'évaluation répond aux deux questions posées plus haut : quel est l'impact des ellipses sur la reconnaissance et la traduction ? Elle montre que l'architecture générale de MedSLT permet une intégration simple avec le contexte qui conduit à une traduction en contexte de qualité suffisante pour la tâche.

2 MedSLT

MedSLT est un système de traduction automatique de la parole (TAP) pour le diagnostic d'urgence de patients étrangers. Il permet à un docteur de poser des questions à un patient dans des domaines spécifiques, comme les maux de tête ou les douleurs abdominales, avec un vocabulaire d'approximativement 1000 à 1500 mots par domaine. En deux mots, MedSLT présente deux spécificités principales (Bouillon *et al.*, 2005). Contrairement aux autres applications de ce type où la reconnaissance est soit statistique (Seligman & Dillinger, 2006; Gao *et al.*, 2006), soit basée sur des grammaires ad hoc de bas niveau (Ehsani *et al.*, 2006), le modèle du langage pour la reconnaissance est compilé ici à partir de grammaires d'unification générales de la langue. Ceci se fait avec la plateforme Regulus, bâtie directement sur les outils vocaux de Nuance(Rayner *et al.*, 2006). Celle-ci permet de dériver toutes les grammaires du système à partir de la même grammaire générale. Pour ce faire, Regulus va d'abord spécialiser les grammaires générales par des méthodes d'apprentissage basées sur des corpus pour les rendre moins ambiguës et les plus spécifiques possibles pour un domaine et/ou une tâche donnés. Il compile ensuite les grammaires spécialisées pour les différentes tâches du système, à savoir : reconnais-

sance/analyse après conversion dans le format CFG requis par Nuance, puis génération. De là, découle la seconde spécificité. Comme ce type de reconnaisance donne surtout des résultats compétitifs pour les phrases couvertes par la grammaire, il s'agit d'une application contrôlée qui suppose que l'utilisateur pourra apprendre la couverture du système, de manière à en tirer le meilleur profit possible. Pour l'aider dans cette tâche, nous utilisons un système d'aide (Starlander *et al.*, 2005; Chatzichrisafis *et al.*, 2006). Au docteur, celui-ci propose, après chaque question, des phrases similaires, couvertes par le système, dont il pourra s'inspirer ; au patient, il donne des exemples de réponses possibles après chaque question. Pour dériver l'aide, le système repose un corpus de phrases préenregistrées avec des questions et les réponses liées. Pour les questions, le système fait en parallèle une reconnaissance statistique et compare ensuite le résultat de cette reconnaissance avec les phrases pré-enregistrées pour extraire les plus similaires (en termes de n-grammes). C'est ainsi que nous introduisons de la robustesse dans un système contrôlé ; pour les réponses, le système d'aide cherche dans le corpus la question la plus similaire à celle qui a été envoyée à la traduction et propose les réponses correspondantes. Une fois reconnue, la phrase est traduite, suivant la méthode interlingue. Dans notre approche, la même grammaire spécialisée est utilisée pour la reconnaissance et l'analyse, ce qui assure que chaque phrase reconnue recevra aussi une analyse. Regulus permet différents types de représentation (Rayner *et al.*, 2006), mais nous avons choisi ici d'exploiter la plus simple possible. Il s'agit d'une structure sémantique quasi plate, formée par concaténation de la sémantique des mots. Par exemple, "avez-vous mal à la tête quand vous buvez du café ?" sera représenté de la manière suivante :

```
[[sc, quand],
 [clause, [[pronoun, vous], [voice, active],
           [tense, present], [action, boire], [cause, café]]],
 [path_proc, avoir], [pronoun, vous], [symptom, mal],
 [tense, present], [utterance_type, sentence],
 [voice, active], [locative_prep, à], [body_part, tête]]
```

Le même formalisme est aussi utilisé pour l'interlangue, qui est une version standardisée de la langue source. Nous avons déjà montré l'avantage de ces structures pour la traduction (Rayner *et al.*, 2005) et comment il est possible de dériver des grammaires spécialisées pour la génération qui évitent le problème de surgénération (Bouillon *et al.*, 2006). Dans la suite, nous verrons qu'elles constituent aussi un excellent point de départ pour le traitement des ellipses.

3 Le traitement des ellipses dans MedSLT

Dans MedSLT, les ellipses sont très productives. On y trouve non seulement des omissions, comme en (7), mais aussi des substitutions (Cf. (8)), où l'ellipse remplace une information déjà existante.

(7) M : avez vous souvent des maux de tête ?
Trad : do you often have the headaches ?
M : **le soir**
Trad : do you often have the headaches **in the evening**

(8) M : avez vous des maux de tête chaque jour ?
Trad : do you have the headaches every day ?
M : **plusieurs fois par jour**
Trad : do you have the headaches **several times a day** ?

Différents types de substitution sont également possibles. Il est par exemple très courant de substituer deux éléments non interrogatifs au niveau de la question, comme en (8). Un élément non interrogatif peut aussi remplacer un élément interrogatif, au niveau de la question (9) ou de la réponse (10). L'algorithme de résolution devra donc être assez puissant pour traiter ces différents cas.

(9) M : when did you visit the doctor ?
Trad : cuándo ha consultado un médico ?
M : **yesterday**
Trad : ha consultado un médico **ayer** ?

(10) M : when did you visit the doctor ?
Trad : cuándo ha consultado un médico ?
P : **ayer**
Trad : i visited the doctor **yesterday**

Comme nous l'avons dit plus haut, nous ne nous appuyons pourtant pas ici sur une analyse profonde. Les représentations plates de MedSLT vont en effet nous permettre un traitement plus efficace et portable, directement adapté des systèmes traditionnels de dialogue homme-machine : comme les phrases elliptiques et leur contexte sont représentés dans des listes plates d'attributs-valeurs, nous pouvons considérer la résolution comme une forme de manipulation de listes d'attributs. Nous traitons ainsi l'**omission** comme une simple concaténation de listes et la **substitution** comme une substitution de listes d'éléments similaires. Pour ce faire, il nous faut cependant répondre à deux questions spécifiques à la TAP : (1) à quel niveau du flux de traitement devons-nous manipuler les listes ? et (2) surtout : comment déterminer les éléments substituables les uns aux autres dans la phrase elliptique et la phrase complète, de la manière la moins ad-hoc possible ?

La Figure 1 illustre le flux de traduction par interlangue dans MedSLT pour les deux phrases *does the pain radiate to your neck ?* et *the jaw ?* Pour résoudre la phrase elliptique et la traduire ensuite en contexte, il faut arriver à remplacer la sous-représentation `[body_part, neck]` de la phrase complète (taggée `[utterance_type, ynq]`), par `[body_part, jaw]` qui se trouve dans la phrase elliptique (notée, `[utterance_type, phrase]`). Dans cet exemple, il serait sans doute préférable de faire la substitution au niveau de l'interlangue. Cette dernière a l'avantage de ne plus contenir l'information sur le possessif (`[possessive, your]`) — les déterminants sont en général tellement mal reconnus que nous avons décidé de ne pas les faire figurer dans l'interlangue et de laisser au générateur le soin de choisir l'article le plus approprié sur la base de corpus (Cf. (Bouillon *et al.*, 2006)). Mais dans beaucoup de cas, la résolution au niveau de l'interlangue posera de nombreux problèmes, notamment quand elle ajoute des informations sémantiques. Par exemple, la phrase "la douleur est-elle aggravée par le chocolat" sera représentée au niveau de l'interlangue dans un sens paraphrasable par "la douleur augmente-t-elle quand vous mangez du chocolat", ce qui rend ensuite la substitution impossible si l'ellipse fait appel à un autre verbe implicite. Nous proposons donc de faire la résolution avant le passage à l'interlangue, ce qui rend d'ailleurs compte du fait que l'ellipse est avant tout un phénomème syntaxique. Celle-ci se fait ainsi en deux étapes. D'abord, nous simplifions le résultat de l'analyse, en enlevant certains éléments, comme le possessif, dont nous

```
Source1 = "does the pain radiate to your neck"
Source Rep1 = [[body_part,neck], [possessive, your],
               [prep,to_loc], [secondary_symptom,pain],
               [state,radiate], [tense,present],
               [utterance_type,ynq]]
Interling1 = [[body_part,neck], [state,radiate], [symptom,pain],
              [tense,present], [utterance_type,ynq]]
Target Rep1 = [[body_part,nuque], [path_proc,irradier],
               [symptom,douleur], [tense,present],
               [utterance_type,sentence]]
Target1 = "la douleur irradie-t-elle la nuque"
Source2 = "the jaw"
Source Rep2 = [[body_part,jaw], [utterance_type,phrase]]
Interling2 = [[body_part,jaw], [utterance_type,phrase]]
Resolved2 = [[body_part,jaw], [state,radiate], [symptom,pain],
             [tense,present], [utterance_type,ynq]]
Target2 = "la douleur irradie-t-elle les machoires"
```

FIG. 1 – Flux de traduction dans MedSLT

venons de parler. C'est ensuite au niveau de cette seconde structure que se font les opérations d'ajout et de substitution. Pour déterminer les éléments substituables, nous exploitons le fait que les mots sont liés dans leur représentation à leur type ontologique (`body_part`, `symptom`, etc.) : un élément peut dès lors se substituer à un autre s'il a le même type ou un type lié (hyponyme, hyperonyme, etc.). Pour appréhender ces éléments de la manière la plus robuste possible et indépendamment des changements dans le lexique ou la grammaire, nous utilisons une méthode d'apprentissage élémentaire. Nous constituons d'abord un corpus qui décrit chaque type d'ellipse possible (lieu, temps, etc.) avec des exemples, comme dans la Figure 2. Ceux-ci sont ensuite analysés par la grammaire, puis généralisés, de manière à extraire les patrons les plus généraux possibles qui ensemble définissent la bonne formation d'une classe d'ellipses. Ici, par exemple, sont généralisés les patrons repris dans la Figure 3. On voit que, en général, seul le type ontologique est gardé, mais pour les éléments interrogatifs, comme *where* ou *when*, nous conservons aussi la valeur de l'attribut de manière à pouvoir distinguer ces mots entre eux.

```
ellipsis_class(
   place_of_pain_np,
   ['the left angle of the ribs', 'the right lower chest',
    'the right chest', 'the front of the head',
    'both sides of the head',
    'both arms', 'the left side',
    'the chest', 'the side']).
```

FIG. 2 – Classe d'ellipses

Sur la base de ces informations, l'algorithme de résolution peut être très simple. Pour toute phrase identifiée comme elliptique par l'analyseur, le système va **d'abord** vérifier si une substitution est possible : chaque sous-représentation d'une phrase elliptique qui correspond à l'un des patrons peut se substituer à toute autre sous-représentation d'une phrase complète qui correspond à un patron de la même classe. Au cas où différentes substitutions sont possibles, le

```
compiled_ellipsis_class(
   place_of_pain_np,
   [[[adj, _], [body_part, _], [part, _]],
    [[adj, _], [adj, _], [body_part, _]],
    [[adj, _], [body_part, _]], [[body_part, _], [part, _]],
    [[body_part, _], [part, _], [spec, _]],
    [[body_part, _], [spec, _]], [[adj, _], [part, _]],
    [[body_part, _]], [[part, _]]]).
```

FIG. 3 – Généralisation des patrons à partir des exemples de la Figure 2

système choisit toujours la plus longue. **Ensuite**, si la substitution est impossible, le système considérera qu'il s'agit d'un ajout. Si un patron manque, le système fera donc erronément un ajout, mais nous verrons dans l'évaluation que ce cas se produit très rarement : la méthode est suffisamment robuste pour apprendre *a priori* des patrons assez généraux pour couvrir les domaines contrôlés de notre application.

4 Evaluation

Dans un premier temps, nous avons voulu vérifier l'impact des ellipses sur la reconnaissance. Nous avons constitué deux ensembles de questions couvertes par le système, avec les questions oui-non types pour les maux de tête. Le premier ensemble ne contient que des phrases complètes (11) ; dans le second, nous avons remplacé les phrases complètes par des phrases elliptiques, quand c'est possible (12). Chaque fichier contient 111 phrases.

(11) avez vous mal quand vous toussez ?
avez vous mal quand vous mangez ?
avez vous mal quand vous êtes stressé ?
avez vous mal quand vous êtes fatigué ?

(12) avez vous mal quand vous toussez ?
quand vous mangez ?
quand vous êtes stressé ?
fatigué ?

Nous avons ensuite demandé à sept personnes de lire successivement les deux séries de questions, puis nous avons mesuré la qualité de la reconnaissance. Pour ce faire, nous avons comparé le taux d'erreurs au niveau des mots (*WER*) et des phrases (*SER*). Comme nous savons ces mesures peu fiables (Wang *et al.*, 2003; Bouillon *et al.*, 2006), nous avons aussi calculé le taux d'erreurs sémantiques (*SemER*), c'est-à-dire le nombre de phrases où la reconnaissance ne préserve pas le sens. Elles correspondent aux phrases dont la rétro-traduction est correcte et qui seraient donc envoyées à la traduction par l'utilisateur. Les résultats sont repris dans le tableau 1. Ceux-ci nous semblent particulièrement intéressants. Les mesures traditionnelles (*WER* et *SER*) donnent l'impression que l'impact des ellipses est très négatif. Elles détérioreraient considérablement la reconnaissance : 11,5% d'erreurs au niveau des mots pour les ellipses versus 3,1% pour les phrases complètes. Mais si on regarde le *SemER*, les résultats ont tendance à s'inverser, dans le sens attendu : le taux d'erreurs sémantiques est **légèrement moins important** pour les

	Ellipses				Phrases complètes			
Sujets	WER	SER	SemER	Secs	WER	SER	SemER	Secs
AR	8,3%	22,4%	0,0%	2,0	1,8%	12,0%	1,3%	3,2
BR	14,3%	32,1%	3,8%	1,6	5,4%	27,3%	3,9%	2,5
CL	13,3%	28,2%	2,6%	2,0	2,0%	11,5%	1,3%	3,0
LA	13,8%	37,2%	0,0%	1,7	4,3%	21,8%	1,3%	2,6
LU	9,5%	25,6%	1,3%	1,8	2,7%	12,8%	3,8%	2,6
PI	10,0%	26,9%	2,6%	2,0	3,5%	21,8%	7,7%	2,8
SE	11,0%	26,9%	0,0%	2,0	2,3%	15,4%	0,0%	3,1
Moyen	11,5%	28,5%	1,5%	1,9	3,1%	17,5%	2,8%	2,8

TAB. 1 – Résultats de la reconnaissance avec des phrases complètes et elliptiques

ellipses que pour les phrases complètes (1,5% versus 2,8%) et ceci pour tous les locuteurs. Un examen des données explique cette différence entre le *WER* et *SemER* : la plupart des fautes faites lors de la reconnaissance des ellipses ne sont pas importantes pour la tâche. Il s'agit par exemple d'erreurs au niveau du nombre : le système reconnaît "occipitales" à la place "occipal", faute de contexte. Mais comme cette information n'apparaît pas dans la représentation, cette erreur n'a aucune répercussion sur la traduction.

Nous pouvons donc tirer trois conclusions de cette évaluation : (1) Les mesures traditionnelles doivent être évitées, puisqu'elles peuvent conduire à de fausses conclusions ; (2) la longueur des phrases n'aurait finalement qu'une légère influence sur la reconnaissance (en tout cas quand cette dernière est basée sur des grammaires et qu'il s'agit de données lues). (3) L'impact des ellipses sur la reconnaissance est donc globalement positif, puisqu'elles diminuent aussi le temps de traitement (*Secs*) : 1,9 secondes de moyenne pour les ellipses versus 2,8 pour les phrases complètes. Mais peut-on obtenir une traduction en contexte de qualité suffisante avec l'algorithme présenté plus haut ?

Pour mesurer la qualité de l'algorithme de résolution, nous avons collecté pour la première fois des données avec le système bidirectionnel anglais-espagnol et nous avons évalué la qualité de la traduction des réponses elliptiques du patient en anglais. Dans cette expérience, l'un d'entre nous a joué le rôle du médecin. Le but était de voir si le patient souffrait ou pas d'une infection bactérienne à la gorge. Les patients étaient au nombre de six. Ils n'avaient jamais utilisé le système auparavant et avaient reçu une carte décrivant leurs symptômes. Nous avons ainsi collecté 47 réponses elliptiques en espagnol, considérées comme bien reconnues par le patient, sur un total de 92 réponses. Celles-ci sont peu nombreuses, mais sont pourtant représentatives : comme le système est contrôlé et dans un domaine très limité, nous n'aurions probablement pas obtenu plus de réponses différentes bien reconnues, en augmentant le nombre de patients. Parmi les réponses, 36/47 ont été jugées tout-à-fait intelligibles et fidèles par deux traducteurs et 3/47, OK. Une seule traduction (1/47) a échoué en raison de l'algorithme de résolution (Cf. exemple (11)), ce qui montre que les patrons appris sont assez généraux pour la tâche. Les autres erreurs se situent soit au niveau de la traduction (2/47), soit au niveau de l'interaction avec le système, par exemple le patient accepte parfois des phrases mal reconnues, ce qui donne des résultats erronés (3/47).

(10) M : for how long have you had your sore throat ?
P : desde hace más de una semana
Trad : i have had a sore throat for more than one week
M : what were the results ?
P : negativo
Trad : the results were negative
M : have you seen a doctor for your sore throat ?
P : sí el lunes
Trad : i visited the doctor for my sore throat monday

(11) M : have you been with anyone recently who has a strep throat ?
P : si más de dos semanas
Trad : i was in contact with someone more than two weeks recently who had strep throat

Une évaluation plus importante est en cours qui devrait compléter ces résultats. Il serait en effet intéressant de vérifier si ceux-ci restent les mêmes avec la version statistique du système et pour des données non-lues. Nous voudrions aussi établir quantitativement qu'une traduction en contexte améliore la tâche.

5 Conclusion

Dans cet article, nous avons montré qu'il est possible d'arriver à traiter adéquatement les ellipses dans un système de TAP, basé sur les grammaires. Comme celles-ci sont très naturelles dans le domaine et semblent avoir un impact positif sur la reconnaissance, il est en effet essentiel de pouvoir les traduire correctement en contexte pour en tirer le meilleur parti possible dans l'application. Pour ce faire, l'approche contrôlée de MedSLT rend possible l'utilisation d'algorithmes de résolution très peu coûteux qui permettent une traduction en contexte intelligible et fidèle. Nous avons ainsi clairement contribué à intégrer le contexte dans une architecture comme la nôtre et, de manière plus générale, dans les systèmes de TAP.

Références

BOITET C., BLANCHON H. & GUILBAUD J.-P. (2002). A way to integrate context processing in the MT component of spoken, task-oriented translation systems. In *Proc. MSC2000*, p. 83–87, Kyoto, Japan.

BOUILLON P., RAYNER M., CHATZICHRISAFIS N., HOCKEY B., SANTAHOLMA M., STARLANDER M., NAKAO Y., KANZAKI K. & ISAHARA H. (2005). A generic multi-lingual open source platform for limited-domain medical speech translation. In *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*, p. 50–58, Budapest, Hungary.

BOUILLON P., RAYNER M., VALL B. N., NAKAO Y., SANTAHOLMA M., STARLANDER M. & CHATZICHRISAFIS N. (2006). Une grammaire multilingue partagée pour la traduction automatique de la parole. In *Proceedings of TALN 2006*, Leuwen, Belgium.

CHATZICHRISAFIS N., BOUILLON P., RAYNER M., SANTAHOLMA M., STARLANDER M. & HOCKEY B. (2006). Evaluating task performance for a unidirectional controlled language

medical speech translation system. In *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, p. 9–16, New York.

EHSANI F., KINZEY J., MASTER D., LESEA K. & PARK H. (2006). Speech to speech translation for medical triage in Korean. In *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, p. 17–23, New York.

FERNANDEZ R. & GINZBURG J. (2002). Non-sentential utterances in dialogue : a corpus study. In *Proc. Third SIGdial Workshop on Discourse and Dialogue*, Philadelphia, PA.

GAO Y., ZHOU B., SARIKAYA R., AFIFY M., KUO H.-K., ZHU W.-Z., DENG Y., PROSSER C., ZHANG W. & BESACIER L. (2006). IBM MASTOR SYSTEM : Multilingual automatic speech-to-speech translator. In *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, p. 57–60, New York.

KIPP M., ALEXANDERSSON J., ENGEL R. & REITHINGER N. (2000). Dialog processing. In W. WAHLSTER, Ed., *Verbmobil : Foundations of Speech-to-Speech Translation*.

RAYNER M., BOUILLON P., SANTAHOLMA M. & NAKAO Y. (2005). Representational and architectural issues in a limited-domain medical speech translator. In *Proceedings of TALN 2005*, p. 163–172, Dourdan, France.

RAYNER M., HOCKEY B. & BOUILLON P. (2006). *Putting Linguistics into Speech Recognition : The Regulus Grammar Compiler*. Chicago : CSLI Press.

SELIGMAN M. & DILLINGER M. (2006). Usability issues in an interactive speech-to-speech translation system for healthcare. In *Proceedings of the HLT-NAACL International Workshop on Medical Speech Translation*, p. 1–8, New York.

STARLANDER M., BOUILLON P., CHATZICHRISAFIS N., SANTAHOLMA M., RAYNER M., HOCKEY B., ISAHARA H., KANZAKI K. & NAKAO Y. (2005). Practising controlled language through a help system integrated into the medical speech translation system (MedSLT). In *Proceedings of MT Summit X*, Phuket, Thailand.

WANG Y.-Y., ACERO A. & CHELBA C. (2003). Is Word Error Rate a good indicator for spoken language understanding accuracy. In *Proceedings of Eurospeech 2003*, p. 609–612, Geneva, Switzerland.

WARD W. & ISSAR S. (1994). Recent improvements in the CMU ATIS system. In *Proceedings of the ARPA Human System Technology Workshop*, p. 213–216.

Analyse automatique de sondages téléphoniques d'opinion*

Nathalie CAMELIN[1], Frédéric BÉCHET[1]
Géraldine DAMNATI[2], Renato DE MORI[1]
[1] LIA/CNRS, University of Avignon,
BP1228, 84911 Avignon cedex 09 France
[2] France Télécom R&D – TECH/SSTP/RVA, 2 av. Pierre Marzin,
22307 Lannion Cedex 07, France
{nathalie.camelin,frederic.bechet,renato.demori}@
univ-avignon.fr
geraldine.damnati@francetelecom.com

Résumé. Cette étude présente la problématique de l'analyse automatique de sondages téléphoniques d'opinion. Cette analyse se fait en deux étapes : tout d'abord extraire des messages oraux les expressions subjectives relatives aux opinions de utilisateurs sur une dimension particulière (efficacité, accueil, etc.) ; puis sélectionner les messages *fiables*, selon un ensemble de mesures de confiance, et estimer la distribution des diverses opinions sur le corpus de test. Le but est d'estimer une distribution aussi proche que possible de la distribution de référence. Cette étude est menée sur un corpus de messages provenant de vrais utilisateurs fournis par France Télécom R&D.

Abstract. This paper introduces the context of the automatic analysis of opinion telephone surveys. This analysis is done by means of two stages : firstly the subjective expressions, related to the expression of an opinion on a particular dimension (efficiency, courtesy, ...), are extracted from the audio messages ; secondly the *reliable* messages, according to a set of confidence measures, are selected and the distribution of the positive and negative opinions in these messages is estimated. The goal is to obtain a distribution as close as possible to the reference one. This study is carried on a telephone survey corpus, provided by France Télécom R&D, obtained in real field conditions.

Mots-clés : détection d'opinions, classification automatique, reconnaissance automatique de la parole, champs conditionnels aléatoires.

Keywords: opinion extraction, automatic classification, automatic speech recognition, conditional random fields.

1 Introduction

Face à la quantité grandissante de donnée disponible, l'extraction d'information pertinente est devenue un des défis de ces dernières années. Plus précisément, l'extraction d'opinion a ré-

*Travaux réalisés en collaboration avec France Télécom's R&D, contrat 021B178.

cemment fait l'objet d'une grande attention de la part de la communauté TALN (atelier d'ACL 2006 *Sentiment and Subjectivity in Text*, DEFT'07). Ce domaine a donné lieu à de nombreuses publications (Wiebe *et al.*, 2005; Choi *et al.*, 2005) portant principalement sur deux aspects : la détection automatique d'opinions à partir d'avis rédigés par des consommateurs (Popescu & Etzioni, 2005) et d'autre part l'analyse de la subjectivité d'une phrase pour les systèmes de résumé automatique ou de question/réponse (Riloff & Wiebe, 2003).

Les travaux présentés dans cette étude concernent le premier cadre applicatif : la détection automatique d'opinions à partir de sondage d'utilisateurs. Il s'agit ici de sondages téléphoniques effectués par France Télécom auprès d'utilisateurs réels. Une des principales caractéristiques de cette étude est la détection d'opinions à partir de messages vocaux, contenant de la parole complètement spontanée, collectée dans des conditions réelles.

A cause des difficultés intrinsèques à ce type de corpus (bruits, téléphone mobile, disfluences, grande variété d'accents, nombreuses digressions entraînant un nombre important de mots hors-vocabulaire), il est très important de développer des méthodes robustes, peu sensibles aux erreurs de Reconnaissance Automatique de la Parole (RAP). En contrepartie, le nombre d'opinions susceptibles d'être détectées est forcément réduit. Nous avons présenté dans (Camelin *et al.*, 2006) une stratégie de RAP utilisant des modèles de langage spécifiques à la détection d'opinions afin de limiter le bruit généré par les portions de messages hors-sujet (ou digressions). La problématique des sondages d'opinions à partir de corpus de messages vocaux et une stratégie d'analyse automatique de ces mêmes sondages a été présentée dans (Béchet *et al.*, 2006). Les travaux présentés ici proposent de nouvelles stratégies mettant l'accent sur l'introduction de connaissances explicites dans le processus de classification automatique. La définition de critères de calibrage du système de détection, spécifiques à la problématique des sondages, est également abordée.

Cet article est organisé comme suit : le paragraphe 2 formalise le problème de la détection d'opinion et de l'analyse de sondages ; le paragraphe 3 introduit le corpus utilisé dans cette étude ; le paragraphe 4 présente les stratégies de détection d'opinions sur des transcriptions manuelles des messages et sur des sorties du module de RAP ; enfin la section 5 permet de comparer ces 2 stratégies et détaille l'ajout de connaissances explicites dans la stratégie analysant les messages vocaux ; elle expose également comment le système peut être paramétré pour le problème particulier de l'analyse de sondages.

2 Formulation du problème

L'analyse automatique d'opinions à partir de messages collectés dans des conditions réelles est une tâche difficile. Une très grande variété de locuteurs exprime leurs opinions de très nombreuses façons, avec des messages de tailles variables, et un grand nombre de répétitions, corrections et contradictions. Á cette variabilité des locuteurs s'ajoute la variabilité acoustique due à des environnements et des canaux de transmissions très variés (téléphones portables, bruits ambiants). Pour toutes ces raisons, les résultats obtenus par les systèmes de RAP sur ce type de corpus sont très variables, avec des taux d'erreurs sur les mots dépassant les 50%.

Ces erreurs de reconnaissance vont grandement affecter les performances de détection d'opinions pour les messages très bruités, cependant le problème de l'analyse de sondages ne nécessite pas le traitement de la totalité des messages enregistrés. En effet, le but recherché est de connaître la distribution des étiquettes d'opinions sur le corpus, pas les étiquettes individuelles

de chaque message. On retrouve ici la problématique traditionnelle des sondages d'opinion : comment collecter un sous-ensemble traitable d'observations qui conserve les mêmes distributions d'opinions que celles du corpus général. Le *sous-ensemble traitable*, dans notre cas, est l'ensemble des messages considérés comme *fiables* par les modèles de RAP et de détection d'opinion. Nous verrons au paragraphe 4 comment cette fiabilité est estimée à l'aide de mesures de confiance.

2.1 Analyse de sondage

Le problème est formalisé de la manière suivante :

Soit C un corpus de n messages oraux $m_1, m_2, \ldots, m_n$.

Soit $C' \in C$ un sous-ensemble de n' messages de C sélectionné par la stratégie d'analyse automatique.

Soit $O(m,x) \in \{o_1, \ldots, o_k\}$ l'opinion exprimé dans le message m à propos de la dimension $x \in D$. Les dimensions correspondent ici aux différentes analyses effectuées dans le sondage, par exemple sur l'efficacité d'un service, la courtoisie des opérateurs, le coût, etc. Les valeurs d'opinions $o_i \in O$ sont les différentes opinions possibles. Dans cette étude on considère les opinions suivantes :

$O = \{$*entièrement positives*, *entièrement négatives*, *mitigées*, *sans opinion*$\}$.

Les opinions $O_r(m,x)$ sont les opinions de référence, données par des annotateurs humains. Les opinions $O_h(m,x)$ sont les étiquettes d'opinions attribuées automatiquement.

Le but des stratégies proposées dans cette étude est de minimiser la distance entre la distribution des opinions de référence PC pour la dimension x :

$$PC(x) = [p(o_1), \ldots, p(o_k)] \quad \textit{et} \quad p(o_i) = \frac{|C_{o_i}|}{n}$$

avec $m \in C_{o_i}$ *ssi* $m \in C$ *et* $O_r(m,x) = o_i$.

et la distribution PC' estimée sur le sous-corpus extrait C' :

$$PC'(x) = [p'(o_1), \ldots, p'(o_k)] \quad \textit{et} \quad p'(o_i) = \frac{|C'_{o_i}|}{n'}$$

avec $m \in C'_{o_i}$ *ssi* $m \in C'$ *et* $O_h(m,x) = o_i$.

Cette distance est évaluée grâce à la divergence de Kullback-Leibler (KLD) entre les deux distributions :

$$D_{KL}(PC(x) \| PC'(x)) = \sum_{i=1}^{k} p(o_i) \cdot \log \frac{p(o_i)}{p'(o_i)} \tag{1}$$

2.2 Détection d'opinions

Le formalisme introduit dans le paragraphe précédent nécessite le calcul de $O(m,x)$ qui représente l'opinion o_i contenue dans le message m à propos de la dimension x. Ces opinions o_i

sont exprimées dans le message sous la forme d'*expressions subjectives* notées $W(o_i, x)$. Par exemple pour la dimension x =*courtoisie* et o_i =*entièrement positive* on peut trouver dans notre corpus : $W(o_i, x) = [$*l'accueil était parfait*$]$.

Le rôle du module de détection d'opinions est d'extraire des messages vocaux m ces séquences $W(o_i, x)$ afin de calculer $O(m, x)$. Cette opération se fait en deux étapes : tout d'abord extraire du message les segments susceptibles de représenter des expressions subjectives, puis caractériser ces segments en fonctions des différentes étiquettes d'opinions.

Une fois ces étapes effectuées, un message m est décrit par une séquence de segments : $m = W_1(o_1, d_1) \dots W_l(o_l, d_l)$ avec $o_i \in O$ et $d_j \in D$.

L'attribution de l'opinion $O(m, x)$ au message m pour la dimension x est effectuée de la manière suivante :

$$O(m,x) = \begin{cases} \textit{sans opinion} & \text{si } \forall\ W_i(o_i, d_i) \in m \text{ on a } d_i \neq x \\ \textit{satisfait} & \text{si } \forall\ W_i(o_i, x) \in m \text{ on a } o_i = \textit{satisfait} \\ \textit{insatisfait} & \text{si } \forall\ W_i(o_i, x) \in m \text{ on a } o_i = \textit{insatisfait} \\ \textit{mitigé} & \text{sinon} \end{cases}$$

3 Description du corpus de sondage téléphoniques

Le corpus de sondage téléphonique a été collecté auprès de réels clients d'un service de France Télécom. Les personnes contactées sont invitées par un court message à appeler un numéro gratuit qui leur permet d'exprimer leur satisfaction vis à vis du service-client qu'ils ont récemment appelé. En composant ce numéro, le message vocal suivant les invite à laisser un message : *[. . .] Vous avez récemment contacté notre service clientèle. Nous souhaitons nous assurer que vous avez été satisfait de l'accueil et de la suite donnée à votre appel. N'hésitez pas à me faire part de tous vos commentaires et de vos suggestions sur notre service, ceux-ci nous aideront à nous améliorer. Nous vous remercions de votre aide et nous restons à votre disposition. Laissez votre message après le signal sonore. [. . .]*

A l'origine ces messages étaient destinés à être traités par des opérateurs. Ainsi aucune consigne de nature à faciliter le traitement automatique n'a été donnée : pas de conseils sur le mode d'élocution, question ouverte et même incitation à laisser des commentaires. Pour cette étude un ensemble de 1779 messages, collectés sur une période de 3 mois, a été transcrit manuellement au niveau mots, expressions subjectives et marqueurs (indication de disfluence et marqueurs discursifs). Ce corpus a été divisé en deux sous-corpus : un corpus d'apprentissage contenant environ 80% des messages et un corpus de test contenant les 20% restant.

L'analyse de la satisfaction des utilisateurs par l'équipe d'analyse des sondages se fait selon trois dimensions : la qualité de l'accueil (notée *accueil*), la rapidité d'accès au service (notée *attente*) et enfin l'efficacité du service (notée *efficacité*). Cette dernière dimension est la plus représenté dans le corpus, elle concerne à la fois l'évaluation des réponses aux attentes des utilisateurs (est ce que le problème a été réglé ?) mais aussi la qualité des informations données. Chaque expression subjective peut recevoir deux polarités : *positive* et *négative*. Nous avons donc un total de 6 étiquettes pour caractériser les expressions subjectives du corpus.

Dans la transcription manuelle, au sein de chaque message, ces expressions sont indiquées par des balises. Nous disposons ainsi d'un corpus de segments, chacun porteur d'une opinion

particulière. Le but du traitement automatique est de retrouver ces segments et de les étiqueter avec l'une des 6 étiquettes. Voici un exemple de message avec les balises manuelles :

oui c'est monsieur NOMS PRENOMS j'avais appelé donc le service client ouais ***<seg label=accueil,pos>*** *j'ai été très bien accueilli* ***</seg>*** *des* ***<seg label=efficacité,pos>*** *bons renseignements* ***</seg>*** *sauf que* ***<seg label=efficacité,neg>*** *ça ne fonctionne toujours pas* ***</seg>*** *donc je sais pas si j'ai fait une mauvaise manipulation ou y a un problème enfin voilà sinon* ***<seg label=efficacité,pos label=accueil,pos>*** *l'accueil était et les conseils très judicieux* ***</seg>*** *même si* ***<seg label=efficacité,neg>*** *le résultat n'est pas n'est pas là* ***</seg>*** *merci au revoir*

4 Détection et classification d'expressions subjectives

Deux stratégies ont été développées pour extraire et classifier les expressions subjectives des messages vocaux : l'une (notée *ref*) s'appuie sur les transcriptions manuelles des messages ; l'autre (notée *asr*) est intégrée dans le processus de décodage de parole. Ces deux stratégies nous permettent de dissocier les erreurs dues à une mauvaise transcription en mots des erreurs de détection d'opinions. Il a été nécessaire de différencier les traitements sur les transcriptions manuelles des traitements sur les transcriptions automatiques à cause de la très mauvaise qualité de ces dernières : les méthodes développées sur le texte *propre* ne sont pas assez robustes pour s'appliquer aux transcriptions automatiques bruitées. La figure 1 présente ces stratégies. Elles sont décrites brièvement dans les prochains paragraphes.

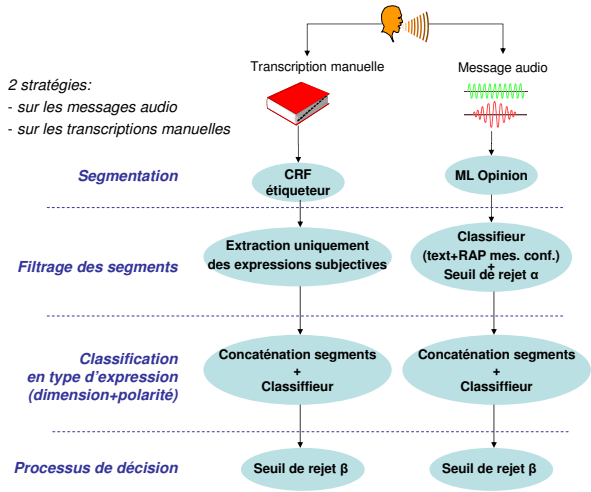


FIG. 1 – Stratégies de détection et de classification d'expressions subjectives pour les transcriptions manuelles et les messages audio

Modèles de Langage spécifiques à l'expression d'opinions. Nous avons montré dans (Camelin *et al.*, 2006) qu'un modèle de langage spécifique à la détection des opinions permettait d'obtenir de meilleures performances lors de la phase de classification qu'un modèle RAP standard de type bigramme. Ceci est dû principalement au protocole de collecte des messages (pas de contraintes d'élocution, encouragement à laisser des commentaires) qui implique une très grande dispersion dans les fréquences de distribution des mots du corpus collecté.

Ce modèle de langage spécifique permet ainsi d'obtenir le message en une suite d'hypothèses d'expressions subjectives séparées par un symbole représentant les segments considérés comme vides.

Un ensemble de mesures de confiance (acoustiques et linguistiques) est associé à chaque hypothèse. Sa probabilité d'être correcte est alors approximée sur le corpus d'apprentissage par régression logistique sur ses mesures de confiance. Comme présenté dans la figure 1 cette probabilité permet de filtrer les hypothèses peu fiables (seuil α de la figure).

Le principal avantage d'un tel modèle est de segmenter directement le flux audio en hypothèses d'expressions subjectives.

Segmentation des messages avec des Champs Conditionnels Aléatoires. Pour le traitement des transcriptions manuelles un segmenteur basé sur les Champs Conditionnels Aléatoires (ou *Conditional Random Fields* CRF) a été développé. Les CRF (Lafferty *et al.*, 2001) ont été utilisés avec succès dans de nombreuses tâches d'étiquetage telles que l'étiquetage morpho-syntaxique ou la détection d'entités nommées. L'avantage principal des CRF par rapport à des modèles génératifs tels que les Modèles de Markov Cachés est la possibilité d'utiliser l'ensemble des observations d'une séquence pour prédire une étiquette. Ce n'est donc pas le seul historique immédiat qui contraint l'attribution d'une étiquette à une observation mais potentiellement toutes les observations précédentes et suivantes.

Dans notre cas, le corpus d'apprentissage est formaté de manière à associer à chaque mot une étiquette indiquant s'il fait partie d'une expression subjective ou non. L'étiqueteur développé est basé sur l'outil *CRF++*[1] qui permet de représenter chaque mot selon différents niveaux. Ainsi un mot est représenté par : son lemme et une étiquette nommée *seed*. En effet, plusieurs études (Hatzivassiloglou & McKeown, 1997) utilisent un ensemble de mots (appelés *seeds*) qui expriment explicitement une opinion (*e.g. gentil*, *agaçant*, *utile*, *efficace*).

Lors de l'analyse d'un nouveau message, les étiquettes posées par le CRF permettent d'extraire uniquement les hypothèses d'expressions subjectives.

Classification des opinions. Une fois la segmentation du message effectuée, les hypothèses d'expressions subjectives associées à un même label sont concaténées. Chaque segment ainsi obtenu est ensuite étiqueté par un classifieur basé sur l'algorithme *AdaBoost* (Schapire & Singer, 2000).

Deux modèles sont appris sur les expressions subjectives du corpus d'entraînement annotées manuellement. La transcription exacte est utilisée pour le modèle *ref* et la transcription automatique pour le modèle *asr*. Chaque expression subjective est représentée en entrée du classifieur par ses lemmes et seeds ainsi que le nombre de mots.

[1] Toolkit CRF++ : http ://www/chasen.org/ taku/software/CRF++/

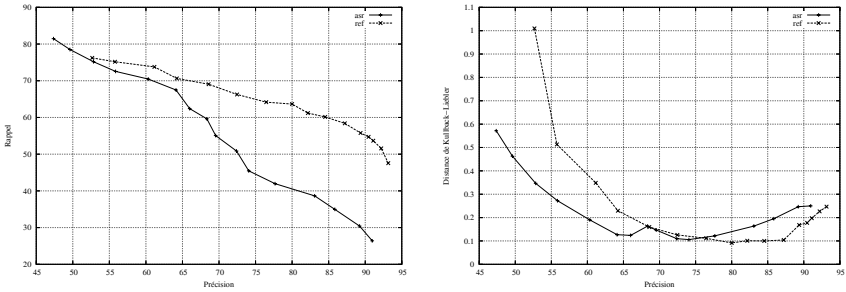


FIG. 2 – Courbes de précision/rappel sur les segments d'opinion (gauche) et de précision/divergence D_{KL} (droite) obtenues sur le corpus de test en faisant varier les seuils de rejet α et β pour les transcriptions manuelles (*ref*) et les sorties du module de RAP (*asr*)

Ainsi l'expression subjective (transcription manuelle) : *j'ai été très satisfaite euh de la communication avec euh l'interlocuteur que j'avais au téléphone donc euh j' attends des documents la personne était très gentille et très serviable et m' a bien dépannée* est représentée par : *il avoir être très satisfaire de le communication avec le interlocuteur que il avoir au téléphone donc il attendre des document le personne être très gentil et très serviable et me avoir bien dépanné , satisfaire communication téléphone attendre personne gentil serviable dépanné , 29*.

En sortie du classifieur, un score de confiance est attribué à chaque étiquette recherchée pour le segment considéré. Les étiquettes retenues sont celles dépassant le seuil β de la stratégie présentée dans la figure 1.

5 Expériences

Cette section permet de comparer les 2 stratégies présentées précédemment et ainsi d'évaluer l'influence des erreurs du module RAP. Une amélioration de la stratégie *asr* avec l'introduction de connaissances explicites est ensuite proposée. Enfin, les critères de choix d'une stratégie par rapport à une autre selon la problématique des sondages sont exposés.

Évaluation des stratégies. L'évaluation est faite sur le corpus de test par rapport à deux types de mesures : les mesures de précision/rappel sur la détection des expressions subjectives et la mesure de la distance de Kullback-Leibler (D_{KL}) entre la distribution de référence sur les opinions et celle estimée automatiquement. La figure 2 présente les courbes obtenues en faisant varier les seuils de rejet α et β.

Comme attendu la stratégie *ref* s'appliquant aux transcriptions manuelles donne de bien meilleurs résultats en terme de précision/rappel. Il est par contre particulièrement intéressant de constater que pour la mesure de la divergence les deux courbes atteignent les mêmes valeurs. Ce résultat valide notre approche consistant à sélectionner un sous-ensemble représentatif de messages pour lesquels les prises de décision sur l'attribution d'opinions sont fiables selon la stratégie implémentée.

Utilisation de connaissances explicites. Une étude manuelle des erreurs générées par la stratégie *asr* sur le corpus d'entraînement a permis de mettre en évidence que de nombreux messages rejetés contenaient des phrases idiomatiques selon une ou plusieurs des dimensions recherchées. Ces phrases ont été extraites du corpus d'entraînement puis manuellement généralisées sous formes d'expressions régulières. Ces expressions sont très générales, en petit nombre, non ambiguë, et sont relativement indépendante de l'application visée de service clientèle. La raison pour laquelle elles n'ont pas été capturées par le processus de classification automatique est due à la faible taille du corpus d'apprentissage. L'apport de connaissance explicite vise ainsi à pallier aux faiblesses des méthodes d'apprentissage automatique sur des données de taille réduite. A la suite de ce processus manuel huit expressions régulière ont été associées à la dimension *accueil*, deux pour *attente* et treize pour *efficacité*.

Pour pouvoir évaluer l'apport de l'utilisation de cette connaissance explicite à notre système, quatre stratégies sont proposées :
- La stratégie Ψ_1 est celle utilisée dans le système *asr*, sans l'apport de connaissances explicites.
- Pour la stratégie Ψ_2, les expressions régulières ont été intégrées comme paramètre d'entrée de l'algorithme de classification *AdaBoost*.
- La stratégie Ψ_3 correspond à la fusion des hypothèses d'opinions obtenues par la stratégie $\Psi 1$ et celles obtenues en appliquant directement les expressions régulières sur les segments.
- Enfin la stratégie Ψ_4 correspond à une stratégie séquentielle : les expressions régulières ayant été apprises principalement sur l'ensemble des messages rejetés par la stratégie Ψ_1, celles-ci sont appliquées uniquement sur l'ensemble des messages rejetés par Ψ_1.

Toutes ces stratégies suivent la règle de rejet suivante : si aucun des segments du message n'a été associé à une étiquette d'opinion, le message est rejeté.

Pour chaque stratégie, la précision, le rappel et la F-mesure ont été calculés en faisant varier les seuils α et β. La figure 3, présentant la F-mesure en fonction de la précision, permet de mettre en évidence l'apport significatif de l'utilisation de connaissances explicites quelles que soit la stratégie choisie, et ce malgré le faible nombre d'expressions régulières rajoutées pour chaque dimension. La stratégie de fusion Ψ_3 est celle qui permet d'obtenir la plus forte valeur de F-mesure.

Choix de la stratégie et réglage du système. Pour les systèmes de détection d'entités, le choix de la meilleure stratégie ou le réglage de paramètres tels que les seuils de rejet est généralement fait sur des courbes de précision/rappel ou de F-mesure telles que la courbe 3. Dans cette étude, le choix de la stratégie à utiliser est fait selon la problématique des sondages d'opinions. En effet, il s'agit de trouver la stratégie qui conservera le mieux les distributions du corpus général. Pour cela la divergence de Kullback-Leibler entre les proportions réelles et celles estimées est calculée pour toutes les stratégies avec différentes valeurs pour les seuils α et β. Le point de fonctionnement du système est choisi comme celui qui minimise cette divergence. La figure 4 présente cette courbe pour les quatre stratégies développées.

Les stratégies Ψ_3 et Ψ_4 montrent une distance de Kullback-Leibler systématiquement plus faible que la stratégie Ψ_1, la stratégie de fusion Ψ_3 apparaissant comme la plus performante.

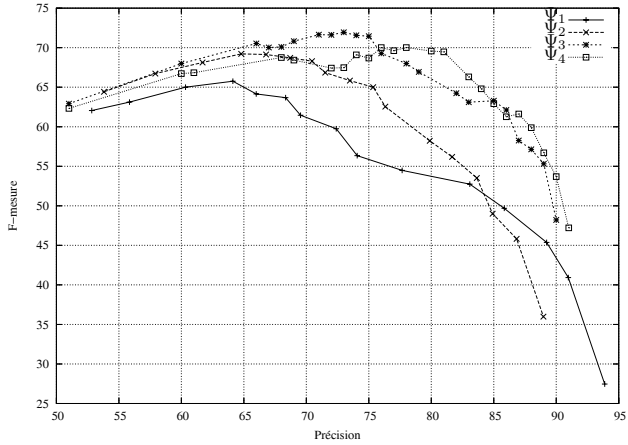


FIG. 3 – F-mesure obtenue par les 4 stratégies d'extraction d'opinions sur le corpus de test en faisant varier les seuils α and β.

6 Conclusion

Nous avons présenté dans cette étude la problématique de l'analyse automatique de sondages d'opinion à partir de messages oraux. Trois résultats originaux ont été obtenus :

1. Il est possible d'extraire de manière robuste de l'information à partir de transcriptions automatiques très bruitées (dues à l'extrême variabilité des corpus oraux collectés) si on accepte de filtrer et sélectionner les messages *fiables* selon un ensemble de mesures de confiance. Les résultats obtenus dans cette étude avec cette stratégie sont identiques à ceux obtenus sur des transcriptions exactes.
2. L'ajout de connaissances explicites peut améliorer un processus de classification automatique en permettant de généraliser certains phénomènes peu représentés dans le corpus d'apprentissage. Diverses stratégies sont proposées pour réaliser cet ajout, c'est la fusion des hypothèses qui s'est montrée la plus robuste dans notre étude.
3. Enfin le choix de la stratégie et son point de fonctionnement doivent être fait par rapport à la tâche visée. Dans le cadre de l'analyse de sondages, c'est la divergence entre la distribution de référence des opinions et celle estimée qui doit être minimisée, plutôt que la précision ou le rappel dans la détection des opinions.

Références

BÉCHET F., DAMNATI G., CAMELIN N. & DE MORI R. (2006). Spoken opinion extraction detecting variations in user satisfaction. In *IEEE/ACL Workshop on Spoken Language Technology*.

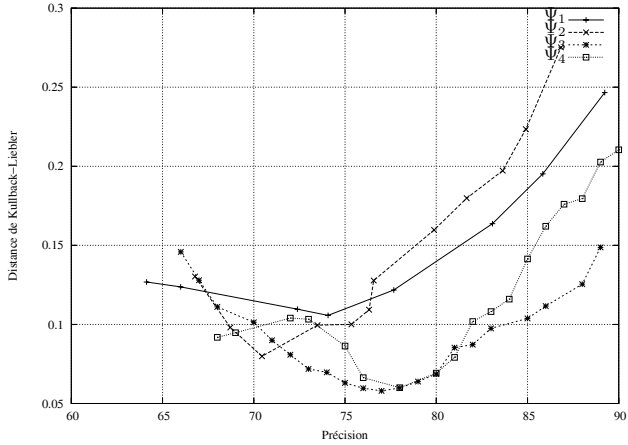


FIG. 4 – Distance de Kullback-Leibler obtenue par les différentes stratégies appliquées sur le corpus de test transcrit automatiquement. Les courbes sont obtenus pour différentes valeurs de α et β.

CAMELIN N., DAMNATI G., BÉCHET F. & DE MORI R. (2006). Détection automatique d'opinions dans des corpus de messages oraux. In *Journées d'Etude de la Parole*, France.

CHOI Y., CARDIE C., RILOFF E. & PATWARDHANN S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *HLT/EMNLP*, p. 355–362, Vancouver, Canada.

HATZIVASSILOGLOU V. & MCKEOWN K. R. (1997). Predicting the semantic orientation of adjectives. In *European chapter of ACL*, p. 174–181, Morristown, NJ, USA : Association for Computational Linguistics.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, p. 282–289 : Morgan Kaufmann, San Francisco, CA.

POPESCU A.-M. & ETZIONI O. (2005). Extracting product features and opinions from reviews. In *HLT/EMNLP*.

RILOFF E. & WIEBE J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on EMNLP*.

SCHAPIRE R. E. & SINGER Y. (2000). BoosTexter : A boosting-based system for text categorization. *Machine Learning*, **39**, 135–168.

WIEBE J., WILSON T. & CARDIE C. (2005). Annotationg expressions of opinions and emotions in language. In *Language Resources and Evaluation*.

Une réalisateur de surface basé sur une grammaire réversible

Claire GARDENT[1], Eric KOW[2]
[1] CNRS/LORIA, Nancy
[2] INRIA/LORIA, Nancy
{Claire.Gardent,Eric.Kow}@loria.fr

Résumé. En génération, un réalisateur de surface a pour fonction de produire, à partir d'une représentation conceptuelle donnée, une phrase grammaticale. Les réalisateur existants soit utilisent une grammaire réversible et des méthodes statistiques pour déterminer parmi l'ensemble des sorties produites la plus plausible ; soit utilisent des grammaires spécialisées pour la génération et des méthodes symboliques pour déterminer la paraphrase la plus appropriée à un contexte de génération donné. Dans cet article, nous présentons GENI, un réalisateur de surface basé sur une grammaire d'arbres adjoints pour le français qui réconcilie les deux approches en combinant une grammaire réversible avec une sélection symbolique des paraphrases.

Abstract. In generation, a surface realiser takes as input a conceptual representation and outputs a grammatical sentence. Existing realisers fall into two camps. Either they are based on a reversible grammar and use statistical filtering to determine among the several outputs the most plausible one. Or they combine a grammar tailored for generation and a symbolic means of choosing the paraphrase most appropriate to a given generation context. In this paper, we present GENI, a surface realiser based on a Tree Adjoining Grammar for French which reconciles both approaches in that (i) the grammar used is réversible and (ii) paraphrase selection is based on symbolic means.

Mots-clés : réalisation de surface, grammaire d'arbres adjoints, réversibilité.

Keywords: surface realisation, tree adjoining grammar, reversibility.

1 Introduction

En génération, le module de *réalisation de surface* a pour fonction de produire, à partir d'une représentation conceptuelle donnée, une phrase grammaticale. Par exemple, à partir de l'entrée donnée en (1a), un réalisateur de surface pourra produire l'une des variantes listée en (1b-1k).

(1) a. *jean(j) aimer(e,j,m) marie(m)*
b. Jean aime Marie
c. Marie est aimée par Jean
d. Marie est aimée de Jean
e. C'est Marie que Jean aime
f. C'est Jean qui aime Marie
g. C'est Marie qui est aimée par Jean
h. C'est Marie qui est aimée de Jean
i. Marie, Jean l'aime
j. Marie, c'est Jean qui l'aime
k. Jean, c'est Marie qu'il aime

Dans cet article, nous présentons un réalisateur de surface (GENI) qui combine une méthode

symbolique de sélection de paraphrases avec une grammaire réversible. En outre, le réalisateur est paramétrable et peut être utilisé soit en mode déterministe (une seule solution produite), soit en mode non déterministe (toutes les paraphrases associées par la grammaire à la sémantique d'entrée sont produites).

L'utilisation de méthodes symboliques pour guider le choix de la paraphrase permet de prendre en compte les facteurs contextuels imposés par un système de génération : la réalisation produite est une réalisation *appropriée* pour un contexte donné plutôt que la réalisation la plus fréquente dans l'usage général représenté par un corpus d'apprentissage.

L'utilisation d'une grammaire réversible a plusieurs avantages.

Premièrement, elle permet d'utiliser le même lexique et la même grammaire pour l'analyse et pour la réalisation. Etant donnée la difficulté de développer de telles ressources, la question de la réutilisabilité est un point non trivial.

Deuxièmement, comme l'illustre la Redwood Lingo Treebank (Velldal & Oepen, 2006), la réversibilité permet de créer rapidement de très grandes suites de tests pour la réalisation : il suffit pour ce faire d'analyser des phrases, de sélectionner parmi les sorties produites l'analyse correspondant à l'interprétation de l'entrée et d'utiliser la représentation sémantique associée comme entrée pour le réalisateur. Par comparaison, les plus grandes suites de tests distribuées actuellement avec les réalisateurs existants soit sont de taille restreinte (500 phrases distinces pour SURGE, 210 pour KPML), soit exigent de développer un module de transformation permettant de créer à partir d'un corpus arboré un format d'entrée adapté à la réalisation (Callaway, 2003).

Troisièmement, la réversibilité permet de mieux mesurer la couverture et le pouvoir paraphrastique du réalisateur. La couverture est testée par une analyse doublée d'une phase de réalisation (la phrase d'entrée peut-elle être à la fois analysée et réalisée par le système ?). Le pouvoir paraphrastique pourra être mesuré en utilisant le réalisateur en mode non déterministe (toutes les paraphrases possibles sont produites) et en identifiant les sorties correctes produites par le réalisateur pour une entrée donnée.

Quatrièmement, une grammaire réversible peut être utilisée à la fois pour la réalisation et pour son inverse à savoir, la construction sémantique (i.e., la construction pour une phrase de sa ou ses représentations sémantique(s)). Si, comme nous cherchons à le garantir dans GENI, le réalisateur a un bon pouvoir paraphrastique, cela a pour effet que la grammaire peut être utilisée à la fois pour générer et pour détecter les paraphrases.

L'article est structuré de la façon suivante. Nous commençons par présenter la grammaire (section 2) et l'algorithme de base (section 3.1). Ce premier algorithme est un algorithme non déterministe qui produit l'ensemble des paraphrases associées par la grammaire à une représentation sémantique donnée. Nous montrons ensuite comment une méthode de paramétrisation des entrées permet de restreindre la réalisation aux seules paraphrases respectant les critères syntaxico-sémantiques spécifiés par les paramètres donnés en entrée (section 3.2). La section 4 présente des résultats chiffrés donnant une indication de la couverture et du pouvoir paraphrastique de GENI et la section 5 compare l'approche proposée avec les travaux connexes. La section 6 conclut avec des pointeurs pour des recherches futures.

2 La grammaire

Formalisme. La grammaire utilisée est une grammaire d'arbres adjoints lexicalisée basée sur l'unification (FLTAG, (Vijay-Shanker & Joshi, 1988)). Une FLTAG comprend un ensemble d'arbres élémentaires et deux opérations permettant de combiner ces arbres entre eux, l'opération de substitution et l'opération d'adjonction. Les arbres résultant d'une de ces opérations sont appelés «arbres dérivés».

Les arbres élémentaires sont lexicalisés, c'est-à-dire qu'ils sont explicitement associés avec un lemme ou une forme fléchie. Leurs noeuds sont étiquetés par deux structures de traits appelées TOP et BOTTOM. Un arbre élémentaire est soit initial, soit auxiliaire. Un arbre initial est un arbre dont les noeuds feuilles sont soit des noeuds terminaux, soit des noeuds dit de substitution (marqués par $\downarrow$). Un arbre auxiliaire est un arbre dont l'un des noeuds feuilles est un noeud «pied» (marqué par $\star$) étiqueté par la même catégorie que le noeud racine.

L'opération de substitution permet d'insérer un arbre élémentaire ou dérivé τ_δ dans un arbre initial τ_α : le noeud racine de τ_δ est alors identifié avec un noeud de substitution dans τ_α et les traits TOP sont unifiés (Top_{τ_α} = Top_{τ_δ}). L'opération d'adjonction permet d'insérer un arbre auxiliaire τ_β dans un arbre quelconque τ_α à un noeud n : les traits TOP_n et BOTTOM_n du noeud n où se fait l'adjonction sont alors unifiés avec les traits TOP du noeud racine de l'arbre auxiliaire et les traits BOTTOM de son noeud pied respectivement (Top_n = $Top_{Root_{\tau_\beta}}$ et $Bottom_n$ = $Bottom_{Foot_{\tau_\beta}}$). En fin de dérivation, les traits TOP et BOTTOM de chaque noeud de l'arbre dérivé produit sont unifiés.

Grammaire et méta-grammaire. La grammaire TAG utilisée est une grammaire produite par compilation à partir d'une spécification plus abstraite appelée, *metagrammaire*. Dans cette métagrammaire, un arbre élémentaire est défini par la combinaison d'un ou de plusieurs fragments d'arbres et chaque fragment d'arbre encapsule une caractérisque linguistique spécifique. Par exemple, tout arbre verbal dont le sujet est un sujet nominal canonique fera intervenir le fragment d'arbre SUJETCANONIQUE. Plus généralement, chaque arbre élémentaire est associé par le processus de compilation à un *profil* listant l'ensemble des noms de fragments d'arbres ayant participé à sa construction. Comme l'illustre le profil de l'arbre donné en Figure 1, le profil de chaque arbre donne ainsi des informations sur ses caractéristiques linguistiques. Or les réalisateurs TAG utilisent souvent ce type d'information pour guider la réalisation (cf. (Yang *et al.*, 1991; Danlos, 1998)) mais l'association est faite de manière manuelle ; Dans GENI, ces informations sont automatiquement associés avec chaque arbre élémentaire par le compilateur de métagrammaire.

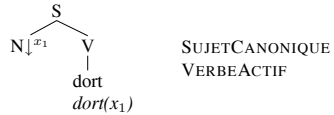


FIG. 1 – Profil d'un arbre

Interface Syntaxe/Sémantique. Dans la grammaire d'arbres adjoints utilisée, le lien entre structure syntaxique et représentation sémantique se fait de la façon illustrée par la figure 2.

Chaque arbre élémentaire est associé avec une représentation sémantique où les arguments manquants sont des variables d'unification. Ces variables apparaissent en outre sur certains noeuds de l'arbre et sont instantiées par le biais des substitutions et des adjonctions[1]. Ainsi, dans la dérivation de *Jean court souvent* illustré ci-dessous, j unifie avec s et r avec x si bien que la représentation sémantique finale est *nom(j,jean), courir(r,j), souvent(r)*.

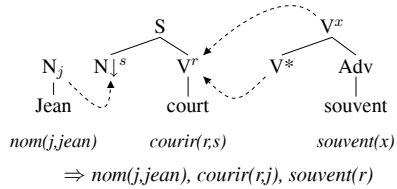


FIG. 2 – "Jean court souvent"

Couverture. La grammaire (Crabbé & Duchier, 2004) couvre la grammaire noyau du français décrite dans (Abeillé, 2002) c'est-à-dire, l'ensemble des cadres syntaxiques de base qui y sont listés et pour chacun de ces cadres, l'ensemble des redistributions (actif, passif, moyen, neutre, réflexivisation, impersonnel, passif-impersonnel) et des réalisations d'arguments permises (cliticisation, extraction, ommission, variations d'ordre).

3 Le réalisateur GENI

Nous commençons par décrire l'algorithme de base utilisé par GENI. Cet algorithme est un algorithme non déterministe permettant de générer l'ensemble des paraphrases associées par la grammaire à une entrée sémantique telle que celle donnée en (1a). Nous montrons ensuite comment cet algorithme peut être modifié pour assurer le déterminisme et générer parmi l'ensemble des paraphrases possibles, la paraphrase respectant les contraintes placées sur une verbalisation par un contexte de génération donné.

3.1 Générer toutes les paraphrases

L'algorithme de base est un algorithme tabulaire (Kay, 1996) ascendant optimisé pour les grammaires d'arbres adjoints. Une spécification détaillée de l'algorithme et des optimisations déployées est donnée dans (Gardent & Kow, 2005). Par manque de place, nous nous contentons ici d'illustrer son fonctionnement par un exemple.

Supposons que la sémantique donnée en entrée soit `courir(r,j),nom(j,jean),souvent(r)`. L'algorithme procède de la façon suivante. Dans un premier temps (**phase de sélection lexicale**), les arbres élémentaires dont la sémantique subsume une partie de l'entrée sont sélectionnés. Pour notre exemple, les arbres sélectionnés seront (entre autre) les arbres de *Jean, court* et *souvent* (cf. Figure 2). La deuxième étape (**phase de substitution**) consiste à explorer systématiquement les possibilités de combinaisons par substitution. Pour l'exemple considéré, cette

[1] Pour plus de détails sur le calcul sémantique utilisé, cf. (Gardent & Kallmeyer, 2003).

exploration permettra de substituer l'arbre pour *Jean* dans l'arbre pour *court* (cf. Figure 2). La troisième étape (**phase d'adjonction**) permet de combiner les arbres produits par adjonction. C'est à ce stade que l'arbre pour *souvent* sera adjoint à l'arbre dérivé pour *Jean court*. En dernier ressort (**phase d'extraction**), les chaînes étiquettant les items couvrant la sémantique donnée en entrée sont produites en l'occurrence : *Jean court souvent*.

3.2 Générer une seule paraphrase

L'algorithme présenté dans la section précédente est non déterministe et produit, pour une entrée sémantique donnée, l'ensemble des paraphrases associées par la grammaire à cette entrée. L'entrée (1a) par exemple, permet la réalisation de l'ensemble des paraphrases listées en (1b-1k).

Pour une utilisation au sein d'un système de génération, il est nécessaire de pouvoir contraindre l'algorithme de facon à pouvoir sélectionner, parmi l'ensemble des paraphrases possibles, la paraphrase appropriée au contexte considéré. Nous montrons maintenant comment l'entrée sémantique de GENI peut être enrichie pour guider l'algorithme dans ses choix et assurer le déterminisme.

Au plus une verbalisation. Dans l'algorithme présenté en section 3, le non déterminisme provient principalement[2] de l'ambiguité lexicale : pour chaque litéral l dans l'entrée, il y a généralement plus d'un arbre élémentaire sélectionné par la phase de sélection lexicale. Ainsi pour chaque entrée sémantique, la sortie de GENI est donnée par l'ensemble des combinaisons d'arbres élémentaires couvrant la sémantique d'entrée et dont la combinaison par substitution ou adjonction est permise par la grammaire.

Pour permettre le déterminisme, nous associons à chaque littéral un *identifiant d'arbre*. Comme nous l'avons vu en section 2, la grammaire utilisée est produite par un processus de compilation qui associe à chaque arbre élémentaire un profil résumant ses caractéristiques linguistiques. Parce que ce profil recense l'ensemble des fragments d'arbres utilisés pour construire un arbre donné et parce que chaque arbre diffère d'un autre par au moins un fragment d'arbre dans son profil, le profil de chaque arbre élémentaire est un identifiant pour cet arbre[3]. Nous utilisons cette caractéristique de la grammaire pour guider la réalisation et assurer son déterminisme de la facon suivante :

1. Chaque litéral l_i dans l'entrée est associée avec un identifiant d'arbre A_i. Cet identifiant est un profil simplifié ne prenant en compte que les informations utiles pour la génération et préservant l'unicité de l'arbre identifié.
2. Pendant la réalisation, pour chaque paire $l_i : A_i$ dans l'entrée enrichie, la sélection lexicale est restreinte aux arbres dont la sémantique subsume l_i *et dont le profil simplifié est* A_i.

Puisque chaque litéral est associé avec un identifiant d'arbre et chaque identifiant d'arbre identifie un arbre unique, le réalisateur produira *au plus* une phrase. Les exemples (2a-2c) illustrent

[2] Une seconde source de non déterminisme provient des modificateurs qui peuvent souvent s'adjoindre dans des ordres différents (*l'homme jeune et grand, l'homme grand et jeune*). Ce type de non déterminisme est traité dans la phase d'adjonction en regroupant les modificateurs d'une même entité et en imposant un ordre unique d'adjonction en cas d'ambiguité.

[3] Il s'agit là d'une simplification. Dans certains cas (peu nombreux), deux arbres élémentaires distincts ont le même profil. Nous revenons sur ce point en section 4.

le type de contraintes mises en jeu par GENI.

(2) a. *l_j :jean(j)/NomPropre l_a :aimer(e,j,m)/[SujetCanoniqueNominal, VerbeActif, ObjetCanoniqueNominal] l_m :marie(m)/NomPropre*
Jean aime Marie
~~Marie est aimée de Jean, C'est Jean qui aime Marie, etc.~~

b. *l_c :le(c)/Det l_c :chien(c)/NomCommun l_d :dormir(e1,c)/SujetRelatif l_r :ronfler(e2,c)/SujetCanoniqueNominal*
Le chien qui dort ronfle
~~Le chien qui ronfle dort~~

c. *l_j :jean(j)/ProperName l_p :promettre(e1,j,m,e2)/[SujetCanoniqueNominal, VerbeActif, ObjetCompletive] l_m :marie(m)/ProperName l_{e2} :partir(e2,j)/VerbeInfinitif*
Jean promet à Marie de partir
~~Jean promet à Marie qu'il partira~~

Au moins une verbalisation. Si une entrée enrichie permet de limiter la réalisation à une verbalisation maximum, elle ne garantit pas une solution au sens où la combinaison d'identifiants d'arbres choisie peut ne pas être réalisable. Mais comment vérifier la satisfiabilité d'une entrée sans pour autant tenter de la réaliser ? Les systèmes existants donnent trois grands types de réponses à cette question.

Une première possibilité est de construire simultanément entrée enrichie et verbalisation. C'est le cas en particulier, des réalisateurs basés sur les grammaires systémiques (KPML(Matthiessen & Bateman, 1991)) où la réalisation coincide avec la traversée d'un réseau systémique permettant d'associer à un contenu sémantique (dimension idéationnelle dans la terminologie systémique), une description syntaxique et fonctionnelle.

Une deuxième possibilité consiste à vérifier la validité de l'entrée sur un critère partiel de bonne formation. Par exemple, REALPRO (Lavoie & Rambow, 1997) prend pour entrée un structure syntaxique profonde de la théorie Sens-Texte (Mel'čuk & Žolkovskij, 1970) et SURGE (Elhadad & Robin, 1999) une description fonctionnelle des grammaires fonctionnelles d'unification. Dans les deux cas, l'entrée n'est pas nécessairement satisfiable puisqu'elle doit être validée par l'ensemble de la théorie. Dans la théorie sens-texte, la structure syntaxique profonde doit, pour être réalisable, pouvoir être projetée successivement en une structure syntactique de surface, une structure morphologique et une structure phonétique. Dans SURGE, l'entrée n'est réalisable que si elle permet une traversée complete de la grammaire du noeud racine aux noeuds lexicaux instanciant la sémantique d'entrée.

Une troisième possibilité consiste à procéder de façon incrémentale et à faire des choix locaux (algorithme gourmand) guidés par les contraintes introduites par les choix faits. C'est le cas en particulier de l'algorithme semi-récursif basé sur les grammaires d'arbres adjoints décrit dans (Danlos, 1998) : à chaque étape de la réalisation, un arbre unique est sélectionné et les contraintes introduites par cet arbre sont utilisées pour contraindre le choix des arbres restant à sélectionner. En cas d'échec, l'algorithme fait un retour arrière.

Nous adoptons une stratégie mixte où la technique dite de *filtrage par polarités* est utilisée pour :

1. effectuer une vérification de la validité de l'entrée enrichie
2. proposer une alternative en cas d'échec de cette vérification

3. proposer une entrée enrichie par défaut en cas d´échec de réalisation

Le filtrage par polarités s'inspire des travaux de (Bonfante *et al.*, 2003; Koller & Striegnitz, 2002) et vise à détecter les combinaisons d'arbres élémentaires où besoins et ressources syntaxico-sémantiques échouent à s'annuler. Plus spécifiquement, l'idée est de vérifier si pour une entrée donnée, l'ensemble d'arbres sélectionnés est tel que *chaque noeud de substitution et chaque noeud pied peut être associé avec exactement un arbre de la catégorie syntaxique et de l'index sémantique approprié.*[4]

Nous utilisons le filtrage par polarité pour filtrer les entrées enrichies non satisfiables (i.e., des entrées dont la polarité est non nulle) mais également pour proposer une entrée alternative en cas de détection d'entrée non satisfiable. Dans ce cas, l'entrée enrichie est dépouillée de ses identifiants d'arbre et l'algorithme non déterministe est utilisé avec additionellement une étape de filtrage par polarité (introduite entre la phase de sélection lexicale et celle de substitution). Les combinaisons d'arbres à polarités neutres (i.e., les combinaisons d'arbres où besoins et ressources syntaxico-sémantiques s'annulent) sont ensuite comparées à l'entrée enrichie initiale et la combinaison la plus similaire à cette entrée est alors proposée en alternative. La similarité entre deux combinaisons d'arbres est mesurée par le nombre d'identifiant communs entre ces deux combinaisons : plus le nombre d'identifiants commun aux deux sélections est grand, plus les sélections sont similaires. Si plusieurs sélections sont également similaires, un choix non déterministe est fait.

Enfin il est malgré tout possible qu'un ensemble d'arbres à polarité neutre soit non satisfiable[5]. Dans ce cas, l'algorithme utilise un enrichissement de la sémantique d'entrée par défaut qui permet de générer une verbalisation «canonique» de cette sémantique.

4 Évaluation

Afin d'évaluer d'une part, le pouvoir paraphrastique du réalisateur et d'autre part, l'impact des annotations de contrôle sur le non-déterminisme, nous avons utilisé une suite de tests graduée. Cette suite a été construite en (i) analysant un ensemble de phrases et (ii) sélectionnant pour chaque phrase la représentation sémantique correcte[6]. Le résultat est une suite de 80 représentations sémantiques choisies pour illustrer les différents types de paraphrases grammaticales décrites par la grammaire utilisée c'est-à-dire,

- les variations grammaticales dans la réalisation des arguments (clivés, cliticisation, extraction, inversion du sujet, etc.) et la forme du verbe (passive/active, impersonnal, etc.)
- les variations dans la réalisation des modificateurs (anté- vs post-posés, adjectif vs subordonnée relative, épithète vs. attribut, etc.)
- les variations permises par une équivalence morpho-dérivationnelle (Ex. arrivée/arriver)

Les 80 cas sélectionnés donnent lieu à la génération par GENI de 1 528 phrases distinctes soit un taux de paraphrases moyen par entrée de 18 avec une variation allant de 1 à plus de 50

[4]Les restrictions d'espace nous empêche d'expliciter ici la méthode de filtrage par polarité. Celle ci est cependant détaillée dans (Gardent & Kow, 2005).

[5]Le filtrage par polarité permet de détecter la non satisfiabilité d'une combinaisons d'arbres pas d'assurer sa satisfiabilité.

[6]L'analyseur peut donner plusieurs analyses et donc souvent plusieurs représentations sémantiques dont certaines représentent correctement le sens de la phrase analysée , d'autres non.

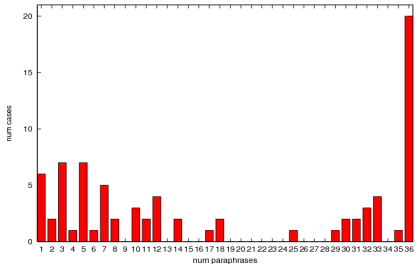


FIG. 3 – Distribution du taux de variation paraphrastique

paraphrases par représentation sémantique. La Figure 3 donne une description plus détaillée de la distribution du taux de variation paraphrastique. Plus généralement, 42% des phrases avec un verbe fini ont une à 3 paraphrases (cas des verbes intransitifs), 44% 4 à 28 paraphrases (verbes prenant deux arguments) et 13% acceptent plus de 30 paraphrases (verbes à trois arguments). Pour les phrases contenant deux verbes finis, le ratio est de 5% des cas ayant 1 à 3 paraphrases, 36% des cas ayant entre 4 et 14 paraphrases et 59% plus de 14 paraphrases. Enfin les phrases contenant plus de 3 verbes finis acceptent toutes plus de 20 paraphrases.

Afin de vérifier que l'utilisation des profils suffit à assurer le déterminisme (cf. section 3.2), nous avons calculé le nombre de cas où deux paraphrases d'un même contenu partagent le même profil. Pour ce faire, nous avons étiqueté de façon automatique les 1 528 paraphrases produites par GENI à partir de la suite de test, avec leur profils (le profil d'une paraphrase est l'ensemble des profils associés aux arbres élémentaires utilisés pour construire l'arbre dérivé de cette paraphrase). Nous avons ensuite comparé, pour chaque entrée de la suite de tests, les profils de toutes les paires de paraphrases correspondantes et compté le nombre de fois où une paire de paraphrases partage le même profil.

Cette manipulation montre que pour les 1 528 paraphrases considérées, le profil échoue à refléter la différence entre deux paraphrases dans moins de 1% des cas. L'analyse des données fautives révèle que les cas posant problème sont les paires paraphrastiques impliquant uniquement (i) une variation d'ordre des arguments (Ex. *Jean donne une pomme à Marie / Jean donne à Marie une pomme*) ou (2) une variation de position pour un modificateur (Ex. *Jean donne ce soir une pomme à Marie /Jean donne une pomme ce soir à Marie / Jean donne une pomme à Marie ce soir*). Le premier cas peut être résolu en modifiant la grammaire de façon à expliciter la différence dans le profil des arbres élémentaires correspondant, le second en imposant un ordre canonique sur l'adjonction des modificateurs.

5 Discussion et comparaison avec travaux connexes

GENI diffère des réalisateurs existants en ce qu'il combine l'utilisation d'une grammaire réversible avec un mécanisme symbolique de sélection des paraphrases.

Ainsi, les réalisateurs utilisant une grammaire réversible utilisent généralement des méthodes statistiques pour choisir parmi l'ensemble des paraphrases associées à un sens par la grammaire,

la paraphrase la plus fréquente plutôt que la paraphrase appropriée à un contexte de génération donné. Ils sont de ce fait mal adaptés à une utilisation dans des scénarios de génération exigeant un bon pouvoir paraphrastique et sont plutôt mis en jeu soit dans des systèmes de traduction (Carroll & Oepen, 2005) soit dans des scénarios de génération où les textes à produire sont fortement stéréotypés et où la paraphrase la plus fréquente est aussi la paraphrase la plus appropriée dans une majorité des cas (White, 2004).

Par ailleurs, les réalisateurs utilisant un mécanisme symbolique de sélection des paraphrases (Danlos, 1998; Matthiessen & Bateman, 1991; Elhadad & Robin, 1999; Lavoie & Rambow, 1997) mettent habituellement en jeu des grammaires qui ne sont pas réversibles c'est-à-dire des grammaires qui ne peuvent pas être exploitées à la fois pour associer un sens à un texte (analyse) et pour verbaliser un sens donné (réalisation). Or comme nous l'avons mentionné dans l'introduction, l'utilisation de la grammaire en analyse comme en réalisation est un point important. Outre qu'elle permet de minimiser l'effort de développement sur la grammaire (chaque modification apportée à la grammaire ou au lexique bénéficie à la fois à l'analyse et à la génération), cette dualité permet un meilleur contrôle de la qualité et de la couverture de la grammaire puisque l'analyseur permet de détecter la sous-génération (i.e., la non génération de phrases grammaticales) et le réalisateur la sur-génération (i.e., la génération de phrases agrammaticales). Enfin, l'utilisation en mode double de la grammaire permet de produire automatiquement des quantités arbitraires d'entrées pour le réalisateur et ainsi de mieux tester son pouvoir paraphrastique. Par comparaison, l'entrée des réalisateurs non réversibles est généralement produite soit par un système de génération, soit manuellement ce qui rend difficile une évaluation précise de leur couverture et pouvoir paraphrastique.

6 Conclusion

En résumé, GENI est un réalisateur de surface basé sur une grammaire d'arbres adjoints pour le français qui présente les propriétés suivantes :

- GENI peut être utilisé en mode déterministe (une paraphrase) ou non-déterministe (toutes les paraphrases)
- GENI utilise une grammaire et un lexique réversible
- GENI utilise une grammaire qui associe aux paraphrases grammaticales une même sémantique – ceci permet, grâce à la réversibilité, à la fois un bon pouvoir paraphrastique et la détection de paraphrases
- GENI est disponible en libre source (`http://trac.loria.fr/~geni`).

Outre les phénomènes décrits en section 4, la grammaire utilisée par GENI est essentiellement la grammaire spécifiée par (Crabbé, 2005) augmentée avec une dimension sémantique (Gardent, 2006). Elle couvre donc l'essentiel de la TSNLP et des phénomènes décrit dans (Abeillé, 2002). Le travail futur inclut (i) un passage à grande échelle par l'extension de la grammaire et du lexique et (ii) l'utilisation du réalisateur pour réduire la sur-génération.

Références

ABEILLÉ A. (2002). *Une grammaire éléctronique du français*. CNRS Editions.

BONFANTE G., GUILLAUME B. & PERRIER G. (2003). Analyse syntaxique électrostatique. *Évolutions en analyse syntaxique, Revue TAL (Traitement Automatique des Langues)*, **44**(3).

CALLAWAY C. B. (2003). Evaluating coverage for large symbolic NLG grammars. In *18th IJCAI*.

CARROLL J. & OEPEN S. (2005). High efficiency realization for a wide-coverage unification grammar. *2nd IJCNLP*.

CRABBÉ B. (2005). *Représentation informatique de grammaires fortement lexicalisées*. PhD thesis, Université Henri Poincaré, Nancy.

CRABBÉ B. & DUCHIER D. (2004). Metagrammar redux. In *CSLP 2004, Copenhagen*.

DANLOS L. (1998). G-TAG : un formalisme lexicalisé pour la génération de textes inspiré de TAG. *Traitement Automatique des Langues - T.A.L.*, **2**.

ELHADAD M. & ROBIN J. (1999). SURGE : a comprehensive plug-in syntactic realization component for text generation. *Computational Linguistics*.

GARDENT C. (2006). Intégration d'une dimension sémantique dans les grammaires d'arbres adjoints. In *TALN 2006*.

GARDENT C. & KALLMEYER L. (2003). Semantic construction in FTAG. In *10th EACL*, Budapest.

GARDENT C. & KOW E. (2005). Generating and selecting grammatical paraphrases. *ENLG*.

KAY M. (1996). Chart Generation. In *34th ACL*, p. 200–204, Santa Cruz, California.

KOLLER A. & STRIEGNITZ K. (2002). Generation as dependency parsing. In *40th ACL*, Philadelphia.

LAVOIE B. & RAMBOW O. (1997). RealPro–a fast, portable sentence realizer. *ANLP'97*.

MATTHIESSEN C. & BATEMAN J. (1991). *Text generation and systemic-functional linguistics : experiences from English and Japanese*. Frances Pinter Publishers and St. Martin's Press.

MEL'ČUK I. & ŽOLKOVSKIJ A. (1970). Towards a functioning meaning-text model of language. *Linguistics*, **57**, 10–47.

VELLDAL E. & OEPEN S. (2006). Statistical ranking in tactical generation. In *EMNLP*, Sydney.

VIJAY-SHANKER K. & JOSHI A. (1988). Feature Structures Based Tree Adjoining Grammars. *12th Computational linguistics*, **55**, v2.

WHITE M. (2004). Reining in CCG chart realization. In *INLG*, p. 182–191.

YANG G., MCKOY K. & VIJAY-SHANKER K. (1991). From functional specification to syntactic structure. *Computational Intelligence*, **7**, 207–219.

Analyse des échecs d'une approche pour traiter les questions définitoires soumises à un système de questions/réponses

Laurent GILLARD, Patrice BELLOT, Marc EL-BÈZE
Laboratoire d'Informatique d'Avignon (LIA)
Université d'Avignon et des Pays de Vaucluse
339 ch. des Meinajaries, BP 1228
F-84911 Avignon Cedex 9 (France)
{laurent.gillard,patrice.bellot,marc.elbeze}@univ-avignon.fr

Résumé. Cet article revient sur le type particulier des questions définitoires étudiées dans le cadre des campagnes d'évaluation des systèmes de Questions/Réponses. Nous présentons l'approche développée suite à notre participation à la campagne EQueR et son évaluation lors de QA@CLEF 2006. La réponse proposée est la plus représentative des expressions présentes en apposition avec l'objet à définir, sa sélection est faite depuis des indices dérivés de ces appositions. Environ 80% de bonnes réponses sont trouvées sur les questions définitoires des volets francophones de CLEF. Les cas d'erreurs rencontrés sont analysés et discutés en détail.

Abstract. This paper proposes an approach to deal with definitional question answering. Our system extracts answers to these questions from appositives appearing closed to the subject to define. Results are presented for CLEF campaigns. Next, failures are discussed.

Mots-clés : système de questions/réponses, questions définitoires.

Keywords: question answering, definitional question answering.

1 Introduction

Les systèmes de Questions/Réponses (sQR) se proposent d'aller au delà de la recherche de documents pertinents afin de répondre, précisément et avec concision, à une question directement formulée en langue naturelle. L'étude de ces systèmes est encouragée par des campagnes d'évaluation qui spécifient des axes de recherche comme la nature des questions à considérer. Cet article s'intéresse tout particulièrement aux questions « définitoires » (QD), en domaine ouvert, telles que *Qui était Alfred Nobel ?* (CLEF06/28), qui interroge sur les aspects biographiques d'un individu, ou les questions *Qu'est-ce que la RKA ?* (CLEF06/95) et *Qu'est-ce que Hubble ?* (CLEF06/02), qui attendent une forme étendue ou encore une (voire *La*) caractéristique remarquable de l'objet à définir. Ces questions ont été introduites lors du volet Questions/Réponses (QR) de la campagne TREC-9. Il est à noter que dans ce qui suit, nous nous limitons au contexte des campagnes EQueR (Ayache *et al.*, 2006) et plus spécifiquement CLEF (Vallin *et al.,* 2006), où une seule et unique réponse est à produire. En effet, depuis 2003 (Voorhees*,* 2003) et dans les campagnes TREC (Voorhees*,* 2005), les réponses attendues aux QD sont constituées de l'intégralité des faits pertinents connus sur le sujet à définir ; cela, au travers d'un découpage en *« pépites »* d'informations vitales (à maximiser), non vitales (indifférentes) et inintéressantes (à minimiser et pénalisantes). *A Contrario*, dans

ce travail, notre objectif est d'extraire *LA* meilleure des réponses pour une QD ; et nous souhaitons y parvenir depuis la détection de mises en apposition et l'emploi d'un minimum de ressources. De plus, ce travail, préliminaire du point de vue de la tâche, nous permet d'explorer l'utilisation de la proximité immédiate des objets à définir avec leur définition. Une autre motivation provenait des faibles performances obtenues par notre système lors d'EQueR sur ces QD : seulement 7% de réponses courtes correctes pour les QD concernant des personnes et 42% pour les autres. Cela est d'autant peu qu'une partie était obtenue grâce à des bases de connaissances et par conséquent la projection de ressources exogènes (couteuses à maintenir). La méthode employée était basée sur un appariement entre un type de réponse attendu et la détection au sein des documents d'Entités du type adéquat. La principale difficulté rencontrée était liée à la l'identification des limites précises de l'énoncé correspondant à la définition. Pour illustrer cette détection parfois mal aisée, considérons une fonction (ou profession), qui débute pourtant par *chef* (une telle construction est assez fréquente en QR) :

- Bouraima Koné, [$_{Fonction}$chef] des opérations techniques] de lutte d'urgence] contre les criquets] au ministère malien] de l'agriculture], au micro de Jean Paul Ade.
(Lu en ligne, http://www2.dw-world.de/french/Politik_Afrika/1.173728.1.html)

Il apparaît, comme le signale les crochets fermants, que la frontière droite de l'expression est délicate à apprécier, et donc sujette à erreurs lors d'un étiquetage en Entités Nommées (ou même une extraction à l'aide de patrons). Par conséquent, le risque existe de faire glisser la réponse extraite vers une réponse incorrecte ou inexacte. Pourtant, cet exemple reste un cas simple : pour les QD qui commencent par *Qui*, la définition à trouver n'est pas systématiquement un rôle social mais peut être n'importe quelle raison pour laquelle une personnalité est connue (de telles QD sont parfois étiquetées *« WhyFamous »*). Et pour les *Qu'est-ce que*, les possibilités sont encore plus vastes. Aussi, notre approche a été de partir des expressions mises en apposition avec les objets à définir pour sélectionner celle qui semble la meilleure (selon différents critères et indices, qui sont présentés en section 3) et de la proposer comme réponse. L'utilisation d'une apposition permet de s'affranchir d'une détection plus hasardeuse, et puisqu'il s'agit d'un extrait du document, cela nous permet de supposer une réponse mieux construite. L'approche que nous présentons est évaluée dans le cadre de notre participation à la campagne QA@CLEF-2006 (Gillard *et al.*, 2006) et obtient autour de 80% de bonnes réponses. Ensuite (section 4), les cas d'échecs que nous avons rencontrés sur CLEF-2006 sont analysés et discutés en détails afin de mettre en évidence des améliorations à mettre en œuvre ou des points importants à considérer pour traiter de telles questions.

2 Autres travaux sur les questions définitoires

Un travail similaire à celui-ci a été fait par (Malaisé *et al.*, 2005) mais il porte sur la détection d'énoncé définitoires dans le domaine médical de la tâche spécialisée de la campagne EQueR. Étrangement, les constructions faisant intervenir des appositions ne sont pas listées dans les nombreux patrons d'extraction lexico-syntaxiques utilisés pour répondre aux questions.

De nombreuses approches ont été envisagées, et pratiquement chaque système emploie une stratégie différente pour traiter les questions définitoires. Par exemple, (Greenwood, Saggion, 2004) utilisent une étape préalable d'acquisition de termes secondaires depuis des ressources exogènes (WordNet, l'encyclopédie Britannica et le Web, ce dernier contribue d'ailleurs a 78%) pour aider à sélectionner des définitions d'abord extraites avec l'aide de patrons. (Fleischman *et al.,* 2003) centrent également leur travaux sur des patrons pour la collecte des réponses candidates, mais se limitent à deux : la succession *Nom Commun Nom Propre* et la

mise en apposition ; ensuite un apprentissage automatique permet un filtrage. L'apposition est également l'un des patrons de (Hildebrandt *et al.,* 2004) parmi 11 autres constructions pour extraire *a priori* des connaissances du corpus ; ils s'aident également d'une projection des définitions de dictionnaires en ligne. (Prager *et al.*, 2001) utilisent les liens d'hyperonymie de WordNet pour localiser un passage contenant une réponse. D'autres, comme (Cui *et al.*, 2005) proposent d'utiliser des patrons lexico-syntaxique probabilistes aux travers de deux modèles (bigrammes et PHMM). (Han *et al.*, 2006) proposent un modèle purement probabiliste basé sur une séparation entre les modèles pour la question et la définition.

3 Une approche et son évaluation pour les questions définitoires

Comme dans tous les systèmes de Questions/Réponses, les questions définitoires sont d'abord identifiées comme telles puis classées suivant quatre catégories (au moyen de motifs d'expressions régulières) : *D+Personne*, pour les questions telles que *Qui est Neil Armstrong ?* (CLEF06/51) ; *D+Acronyme* pour les questions telles que *Qu'est-ce que l'OUA ?* (CLEF06/48) ; *D+minuscules* pour celles comme *Qu'est-ce que l'effet de serre ?* (CLEF06/189) ; et enfin, ce qui constitue le choix par défaut, la catégorie *D* pour les questions définitoires qui n'entrent dans aucune des précédentes catégories comme par exemple *Qu'est-ce que Challenger ?* (CLEF06/81), *Qu'est-ce qu'Euro Disney ?* (CLEF06/29) ou *Qu'appelle-t-on le Knesset ?* (CLEF06/103). Ensuite, l'objet à définir est obtenu en filtrant, après un étiquetage morphosyntaxique (TreeTagger), les différents pronoms interrogatifs et mots vides de sens comme *être*, *appeler*, *acronyme*, *sigle*, *etc.* Puis toutes les phrases du corpus contenant l'objet à définir sont conservées et étiquetées morphosyntaxiquement. Si aucune phrase n'est trouvée la réponse *NIL* est retournée.

Ce n'est qu'après ces différents prétraitements que les différentes expressions candidates sont extraites. Chacune d'entre elles est accompagnée de différents critères qu'il est possible de percevoir comme des juges dont les votes pondérés permettent de faire préférer *in fine* l'une plutôt que l'autre. La fusion de ces différents jugements est variable suivant l'appartenance de la question à l'une des quatre catégories initiales. Les réglages utilisés ont été obtenus de manière empirique sur les QR des campagnes CLEF-2004, 2005 et EQueR.

Les expressions apposées à celle à définir et séparées d'elle par des virgules sont extraites et constituent l'ensemble préférentiel dans lequel la réponse devrait être extraite. Un critère correspondant à leur fréquence d'apparition dans l'ensemble des phrases leur est associé. Un autre ensemble plus particulièrement adapté aux acronymes et abréviations est défini et correspond à une construction *Expression (CAPITALES)* ou *CAPITALES (Expression)*, par l'intermédiaire de deux extractions de ces *Expressions* : l'une est obtenue depuis un alignement dynamique entre la forme *CAPITALES* et l'*Expression*, l'autre depuis la partie gauche (respectivement droite) la plus redondante ; là encore, un critère de fréquence est associé avec chacune d'elles. D'autres juges sont définis : la présence en tête de l'expression du nom le plus fréquent à la position immédiatement à gauche ; de même, et après application d'un motif morphosyntaxique minimal pour détecter des groupes nominaux, la présence en tête de ce groupe nominal du nom principal déterminé le plus fréquent ; un taux de couverture avec le centroïde des noms les plus fréquents au sein d'une fenêtre de 10 mots autour de l'objet à définir ; un fonction de la longueur de l'expression ; un fonction du nombre de noms ; et deux autres juges binaires pour la présence des noms *président* ou *société* en tête d'expression. Enfin, depuis ces expressions, et une stratégie de fusion des provenances, catégories et juges, les meilleures sont proposées comme réponse avec un comportement par défaut qui consiste à répondre par : l'expression apposée qui est à la fois la plus longue et en adéquation avec le motif de détection des groupes nominaux, si elle existe, ou le nom le plus

fréquent précédant l'objet à définir. Cette réponse par défaut est systématiquement proposée en position 5.

Contexte et évaluation de la méthode : Le tableau 1 propose un référentiel pour l'évaluation de notre méthode au travers d'un bilan sur les résultats obtenus par l'ensemble des participants aux questions définitoires en français des 3 campagnes QA@CLEF passées, mais selon les ventilations que nous avons retenues. En 1ière colonne (*#Q*) est présenté le nombre de QD pour chacune des campagnes et catégories : 20 en 2004, 50 en 2005 et 42 en 2006. Le 2ième groupe de colonnes correspond aux nombres et pourcentages de ces questions pour lesquelles une réponse correcte a été trouvée (RC) par au moins l'un des participants. Il ressort de cette colonne que l'ensemble des stratégies mises en œuvre par tous les systèmes permet, à partir de 2005, de s'approcher de la totalité des réponses à obtenir (94%). Enfin le dernier groupe (*RC par soumission*) permet de situer les performances des systèmes puisque figure le nombre de réponses correctes obtenues par : la/les moins bonnes des soumissions (*Min.*), la/les meilleures (*Max.*), ainsi que leur moyenne arithmétique (*Moy.*). Il est à noter que l'année 2004 est moins représentative puisqu'un seul système a participé. Également, si le (ou les) meilleur des systèmes dépasse 80% de bonnes réponses, la moyenne de ceux-ci se situe en deçà (34% et 50%). Les questions portant sur les mots/concepts les plus généraux (*D+minuscules*), et dont la réponse devrait être proche d'une définition du type dictionnaire, rencontrent une réussite moindre, il est possible d'envisager deux raisons à cela : le fait que les corpus journalistiques employées se prêtent peu à ce type d'extraction (contrairement à des d'informations plus biographiques), ou tout simplement leur relative nouveauté en QR.

		#Q	*Réponses Correctes trouvées (RC)*		*RC par soumission*				
					Min.	*Max.*		*Moy.*	
CLEF-2004 *1 participant, 16 soumissions*	D+Personne	12	7	58%	1	6	50%	3	25%
	D+Acronyme	5	0	0%	-	-	-	-	-
	D	3	0	0%	-	-	-	-	-
	Total	***20***	***7***	***35%***	***1***	***6***	***30%***	***3***	***15%***
CLEF-2005 *6 participants, 12 soumissions*	D+Personne	25	24	96%	2	22	88%	11,50	46%
	D+Acronyme	22	21	95%	0	20	91%	10,75	49%
	D	3	2	67%	0	1	33%	0,33	11%
	Total	***50***	***47***	***94%***	***2***	***43***	***86%***	***16,94***	***34%***
CLEF-2006 *7 participants, 15 soumissions*	D+Personne	17	16	94%	1	15	88%	9,67	57%
	D+Acronyme	5	5	100%	0	5	100%	2,33	47%
	D	16	15	94%	1	13	81%	8	50%
	D+minuscules	4	3	75%	0	2	50%	1	25%
	Total	***42***	***39***	***93%***	***2***	***35***	***83%***	***21***	***50%***

Tableau 1: Bilan sur les réponses correctes(RC) proposées par les participants aux questions définitoires des campagnes QA@CLEF.

Le tableau 2 présente les résultats obtenus par notre méthode sur les mêmes jeux de QD que le tableau 1. Les questions des campagnes CLEF 2004, 2005 et EQueR (non présentées) ont été utilisées pour raffiner la méthode mise au point après notre participation à EQueR. L'évaluation de ces résultats a été faite manuellement. Il en est de même pour la dernière colonne (*Au moins une réponse correcte dans les 5 premières réponses*). En revanche pour les autres données de CLEF 2006, les résultats présentés sont ceux obtenus lors de notre participation à la campagne au volet Français-Français (les résultats en Anglais-Français sont inférieurs en raison d'erreurs de traduction ; 67% de réponses correctes sont trouvées au rang 1 au lieu de 79%). Les *(+1)* et *(+2)* qui figurent dans le tableau correspondent à des réponses qui auraient dû être extraites mais qui ont été perdues à cause de problèmes d'ingénierie, ou pour l'une de la ligne *D+Personne* en raison d'une incertitude que nous avons : est-il

acceptable de définir *Boris Becker* comme une *tête de série n°7* ? (*cf.* discussion en section suivante). Il apparaît de ce tableau que les taux de bonnes réponses sont constants et aux alentours de 80% au premier rang et dépassent 83% pour une réponse correcte placée parmi les 5 premières. Cependant, et comme précédemment souligné par le tableau 1, les questions définitoires portant sur des concepts génériques (*D+minuscules*) ou qui ne concerne ni les personnes ni les acronymes, rencontrent également moins de succès avec notre méthode.

		#Q	*Au rang 1, Réponses*			*Au moins une Réponse Correcte dans les 5ières*	
			Correctes (RC)		*ineXactes*		
CLEF-2004	D+Personne	12	9	75%	2	12	100%
	D+Acronyme	5	5	100%	-	5	100%
	D	3	2	67%	1	3	100%
	Total	***20***	***16***	***80%***	***3***	***20***	***100%***
CLEF-2005	D+Personne	25	20	80%	3	21	84%
	D+Acronyme	22	20	91%	2	22	100%
	D	3	1	33%	1	2	67%
	Total	***50***	***41***	***82%***	***6***	***45***	***90%***
CLEF-2006	D+Personne	17	15	88%	1	15 (+2)	88%
	D+Acronyme	5	4 (+1)	80%	-	4	80%
	D	16	13	81%	2	15	94%
	D+minuscules	4	1	25%	-	1	25%
	Total	***42***	***33***	***79%***	***3***	***38***	***83%***

Tableau 2: Résultats obtenus par notre méthode sur les questions définitoires des campagnes QA@CLEF ; évaluation officielle pour 2006 (soumission FR-FR).

4 Analyse et discussion des cas d'échecs sur CLEF-2006

Cette partie propose une discussion sur les cas d'échecs que nous avons rencontrés. Aussi, elle s'articule autour de quelques questions qui illustrent des difficultés représentatives pour notre système et parfois la tâche elle-même. Il faut également rappeler que notre méthode n'utilise pas d'analyse syntaxique et repose essentiellement sur une recherche d'expressions apposées depuis un découpage en phrases, et, par conséquent, avec un contexte limité.

Qui est Boris Becker ? (CLEF2006/90) Au delà de la simplicité évidente de cette question pour un « lecteur moyen de journaux », la difficulté est réelle puisqu'aucun des systèmes n'a trouvé une réponse correcte. En effet, la réponse retournée par notre système est une nationalité au travers du nom *Allemand* soit le nom qui qualifie le plus fréquemment *Boris Becker* dans le corpus (*l'Allemand Boris Becker* est présent 21 fois sur les 104 occurrences de *Boris Becker*). Cette réponse n'a pas été jugée correcte, la raison étant qu'une nationalité n'est pas suffisante, à elle seule, pour qualifier convenablement une personne. Cette règle connue, il apparaît possible, notamment dans ce cas simple, de filtrer les « mauvais » candidats à l'aide de règles de rejet. Aussi, il faut poursuivre l'investigation. L'une des premières définitions qui viendrait à l'esprit pour définir *Boris Becker* serait très probablement sa qualité de *joueur de tennis*. Mais, au sein d'un découpage en phrase du corpus, *Boris Becker* entre que très rarement en cooccurrence avec un motif comprenant *tennis* (qu'il soit issu de l'expression *joueur de tennis* ou même *tennism(a|e)n*). Et, sur les neuf fois où cela se produit sur les 104 phrases contenant *Boris Becker*, une seule permet effectivement de le définir directement au travers de ce trait :

- Sa fortune , il l' a bâtie sur les courts de tennis aux côtés de son partenaire de double , le fantasque Ilie Nastase , dans les années 60 ; puis en gérant les affaires des meilleurs tennismen , tels que l' Allemand Boris Becker hier ou le Croate Goran Ivanisevic aujourd'hui . (LEMONDE94-000881-19940108)

Cependant, malgré la présence du *tels* introductif à une explicitation, il faut souligner ici à quel point le processus de réponse apparaît délicat : il est nécessaire de ne pas tenir compte de

la présence du nom *Allemand* pour revenir en arrière jusqu'à celui de *tennismen*. En outre, une inversion de la position des deux joueurs *Goran Ivanisevic* et *Boris Becker* dans la phrase aurait encore complexifié son analyse et aurait nécessité la mise en balance des deux groupes nominaux par la conjonction *ou*. Et dans ce dernier cas, la tâche aurait été compliquée par la présence des mots *hier* et *aujourd'hui* puisqu'ils apparaissent comme des éléments perturbateurs difficiles à prévoir (notamment dans le cas d'une extraction à partir de patrons morphosyntaxiques) mais pourtant à ignorer. Enfin, il faut se remémorer qu'il s'agit de l'unique cooccurrence entre *tennismen* et *Boris Becker* dans le corpus et par conséquent une prise en compte fréquentielle n'est pas envisageable dans ce cas (cependant une parenthèse mérite d'être ouverte : des procédés de résolutions d'anaphores pourraient augmenter le nombre de candidats mais notre système en est actuellement dépourvu).

Cette (probable) bonne réponse étant écartée, il est possible de s'intéresser à une autre réponse candidate ou plutôt un autre lot de réponses envisageables. En effet, *Boris Becker* est à plusieurs reprises qualifié de *tête de série* comme explicité dans les exemples ci-dessous :

- Victime d' une blessure au dos à l' échauffement , l' Allemand Boris Becker , tête de série numéro 10 , a déclaré forfait avant le match contre Stark (LEMONDE94-002991-19940525)
- Pete Sampras a gagné , dimanche 15 mai , les Internationaux d' Italie de tennis en battant en finale l' Allemand Boris Becker , tête de série n 8 (6-1 , 6-2 , 6-2) . (LEMONDE94-001994-19940517)
- Sur le court n° 1 , dos-à-dos à deux sets partout , l' Allemand Boris Becker , tête de série n° 7 , et l' Ukrainien Andrei Medvedev (n° 9) ont vu leur rencontre interrompue [...] (LEMONDE94-003405-19940629)

Ainsi, une première difficulté survient dans le cas d'une éventuelle factorisation fréquentielle sur *tête de série :* l'expression n'est pas suffisante si le qualifié n'est pas *tête de série numéro 1*. Aussi, il peut être nécessaire de la complémenter. Dans ce cas, cela suppose d'être en mesure de prendre en compte les différentes écritures de *numéro*. Pourtant cela serait une erreur de penser qu'il s'agit d'un même classement comparable et qu'il est possible de suivre sa variation dans le temps au travers des différents documents (auquel cas, une décision aurait pu être de ne considérer que le plus récent). En effet, une *tête de série* correspond à un classement préalable à un tournoi, sorte d'estimation faite en fonction du niveau d'un participant, pour faire en sorte que les meilleurs d'entre eux ne se rencontrent qu'à la fin de la compétition. Aussi la notion de *tête de série* n'a de sens que vis-à-vis d'une compétition sportive. Cependant, comme il est possible de le voir sur ces exemples, ces références ne sont pas toujours présentes dans la fenêtre de la phrase : seul le deuxième exemple propose à la fois la compétition et le domaine au travers des *Internationaux d'Italie de tennis*. Aussi, et finalement, l'ensemble de ces ambiguïtés liées à autant d'élisions peut diminuer considérablement l'intérêt d'une réponse extraite de ces phrases (sauf à pouvoir synthétiser une réponse telle que *tête de série n°8 aux Internationaux d'Italie de tennis* mais dans ce cas il serait probablement préférable d'aller jusqu'à *demi-finaliste aux Internationaux d'Italie de tennis en 1994*).

Il est d'ailleurs à noter qu'une partie de cette discussion eut été différente si plutôt que *tête de série n°X*, l'expression *n°X mondial* avait été présente, puisque alors il aurait fallu prendre justement en compte une variabilité dans le temps de ce classement parmi les meilleurs joueurs mondiaux. Et de s'interroger sur l'opportunité de qualifier *Boris Becker* avec autant de classements différents malgré un dénominateur commun d'être « l'un des 10 premiers joueurs mondiaux de tennis en 1994 » (d'ailleurs, n'est-ce pas la réponse, actuellement hors de portée, qu'il aurait fallu pouvoir inférer de ces différentes réalisations ?).

Après avoir considéré ces candidats de réponses propulsés en tête des possibles en raison de leur répétition, il apparaît parmi les appositions restantes quelques autres susceptibles de donner lieu à des réponses correctes. Cependant, dans trois de celles-ci, le rapprochement

avec le monde du *tennis* n'est pas présent, ce qui par conséquent amènera à présupposer cette connaissance (qui peut ne pas aller de soi). Il est aussi intéressant de noter que trois d'entres elles commencent par des adjectifs numéraux ou multiplicateurs (peut-être faut-il y voir une particularité de l'univers sportif ?). Enfin, et comme pour étayer la discussion passée, une erreur de typographie peut compliquer les rapprochements des *tête de sréie numéro 3*. Tout comme il apparaît délicat de décider si *Boris Becker* est *huitième joueur mondial* ou bien *no 3 mondial* (respectivement dans un document de 1994 et 1995, aussi a-t-il été successivement, l'un puis l'autre ; du moins au moment de l'écriture de chacun de ces documents).

- En déclarant , [...] sans toutefois apporter de preuves , l' Allemand Boris Becker , triple champion de Wimbledon , a relancé la rumeur autour du dopage dans le monde du tennis (LEMONDE94-000256-19940104)
- L' Allemand Boris Becker , huitième joueur mondial et tête de sréie numéro 3 , s' est imposé en finale du tournoi de New-Haven (Connecticut) en battant [...] (LEMONDE94-001895-19940823)
- La fédération accéderait ainsi aux exigences de Stich , qui réclame une somme identique à celle accordée à son compatriote Boris Becker , no 3 mondial , soit 1 million [...]. (LEMONDE95-010221)
- Boris Becker , multiple vainqueur dans les autres tournois du Grand Chelem , a fait une nouvelle fois son deuil des Internationaux de France . (LEMONDE95-022102)

Enfin, il est possible de remarquer que, depuis ce dernier exemple, et puisque la réponse est à fournir hors du contexte du document sans doute serait-il préférable d'être capable de perdre le *autres* pour ne conserver que *multiple vainqueur dans les tournois du Grand Chelem*.

Qui était Alexander Graham Bell ? (CLEF2006/0050) C'est un problème d'ingénierie lié aux nombreuses ponctuations qui a empêché notre système de répondre à cette question depuis :

- Parmi les lieux historiques , le canal de St-Peters (entre le lac du Bras-d' Or et l' Atlantique) , le site dédié à l' inventeur du téléphone , Alexander Graham Bell (à Baddeck , à l' ouest de Sydney) , Louisbourg (lire notre reportage) , la citadelle de Halifax (fortifications du XIXe) , Port-Royal , à 210 km à l' ouest de Halifax [...] . (LEMONDE95-023507)

Mais cette question nous permet d'illustrer une autre difficulté qu'il ne faut pas oublier de considérer lors des étapes de recherche ou d'appariement, qu'il s'agisse de la question, des documents voire de l'un avec l'autre. En effet, il faut autoriser une certaine variabilité dans les noms propres afin de s'assurer qu'*Alexandre* et *Alexander* soit une même personne, qu'un éventuel nom intermédiaire, deuxième prénom, *etc.* puisse également être facultatif ou présent. Cela pour qu'*Arantxa Sanchez Vicario* (CLEF05/175) puisse s'écrire *Arantxa Sanchez-Vicario* mais n'être parfois qu'*Arantxa Sanchez*. S'il en doit en être de même pour *Bill*, *William Jefferson, W.J.* ou même *William J. Clinton*, force est de constater que dans ce cas la solution n'apparaît pas triviale (quelle forme canonique pour les noms propres ?). Enfin, dans certains cas, plus chanceux, si l'écriture fautive apparaît à la fois dans la question et les documents, le système peut établir qu'*Hil(l)ary Clinton* (CLEF06/120) est tout de même *l'épouse du président américain*. Néanmoins, il est aussi possible d'utiliser des algorithmes de tolérance aux fautes.

- Un compromis [...]sans délai en échange de la confirmation par Washington de la présence de Hilary Clinton , l' épouse du président américain , à la conférence [...]. (LEMONDE95-031398)

Qu'est-ce que le Crédit Suisse ? (CLEF2006/0107) Cette question est particulièrement intéressante : comment peut-on définir quelque chose qui apparaît comme déjà transparent ? Notre système n'y est pas parvenu et s'y est même trompé : le *groupe Crédit Suisse* est effectivement un *groupe* (et à de nombreuses reprises), mais ce n'est pas très satisfaisant et si c'est le *seul groupe*, c'est déjà trop. En effet, il faut prendre garde à certains adjectifs qui nuisent plus qu'ils n'apportent. Nous avions envisagé ce problème mais oublié de l'implémenter : s'il est intéressant de savoir que *Stephen Hawking* (CLEF06/0027) est un *célèbre physicien anglais*, ou que

Nick Leeson (CLEF06/0041) est un *ancien courtier de la banque Barings* ; d'autres adjectifs tels *nouveau* doivent être soigneusement évités (et à plus forte raison pour un corpus ancien). Ainsi notre système aurait pu filtrer *Windows* (CLEF06/107), le *nouveau système d' exploitation* jugé inexact (mais aussi parce qu'il était question de l'une des versions de *Windows* dans le document). En outre, et pour revenir au cas *Crédit Suisse*, les systèmes participants n'ont pu faire mieux que de répondre *groupe* qui a été la seule bonne réponse acceptée, à deux reprises, lors de l'évaluation (mais il reste possible de s'interroger sur l'intérêt de cette définition, surtout lorsque *Prix* de *Le Prix Crédit Suisse* est refusé, peut être parce qu'il s'agit d'une *« copie »* ou une *« imitation »*).

Il est à noter que notre système propose en 4ième position, *banque qui a placé les emprunts Biber*, depuis :

- Ces créanciers spéculent [...] hausse des cours , a expliqué Dieter Enkelmann , membre de la direction du Crédit Suisse , banque [est-ce à conserver ? qui a placé les emprunts Biber] . (ATS.940426.0097)

mais la fréquence *banque* (7) n'a pu prendre le pas sur celle de *groupe* (31). Un autre point est qu'il peut apparaître opportun (d'ailleurs plus par conformité avec les spécifications des réponses à produire dans le cadre des campagnes d'évaluations CLEF que du point de vue de la pertinence de la *« pépite »* d'information elle-même) d'éloigner la proposition subordonnée relative pour ne conserver que *banque*. Et encore une fois, dans le cas de *Windows* (CLEF06/107) :

- Microsoft Network [...] que la prochaine version de Windows , le système d' exploitation [est-ce à conserver ? qui a permis au groupe d' asseoir sa domination du marché des logiciels] (ATS.950516.0148)

Pour conclure avec l'institution financière, dans l'exemple ci-dessous, il ne faut pas extraire trop rapidement *banque bâloise* puisqu'il s'agit en fait de la *Société de Banque Suisse (SBS)* et non du *Crédit Suisse*, l'une des *grandes concurrentes* du *SBS*. Mais la compréhension nécessaire n'apparait pas dans la fenêtre de la phrase.

- A l' instar de ses deux grandes concurrentes , l' UBS et le Crédit Suisse , la banque bâloise a subi le contre-coup des turbulences qui ont affecté l' an passé les marchés financiers . (ATS.950315.0084)

Qu'est-ce qu'un samovar ? (CLEF2006/0188) Cette question cristallise les difficultés. Dans le corpus, il n'est pas possible d'obtenir la définition à laquelle on s'attend. Pourtant, c'est justement grâce à une apposition qu'il est envisageable de répondre depuis l'une des 9 phrases utilisant le mot (les 8 autres sont susceptibles de brouiller les pistes). Mais l'analyse pour aboutir apparaît particulièrement complexe : il faut considérer la seconde occurrence du mot plutôt que la première (laquelle seconde est à ignorer sinon), passer outre la forme plurielle, les guillemets fermants, une partie de la première apposition, pour s'arrêter après la troisième, et non plus loin. Ensuite, un samovar peut être *ceux qui, au front, avaient perdu bras ou jambe*. Aucun des systèmes participants n'y est parvenu. Et finalement, cette question pose un autre problème sous-jacent à toute la tâche Questions/Réponses : est-il toujours possible d'extraire automatiquement une réponse concise depuis un passage qui la contient manifestement ?

- Près du [1samovar] , ou entourée de " [2samovars] " , comme on appelait ceux qui , au front , avaient perdu bras ou jambe , la voilà soudain loin de la littérature , se souvenant [...] (LEMONDE95-036764)

Qu'est-ce qu'un t-shirt ? (CLEF2006/0144) Cette question a donné lieu à une réponse vestimentaire sans intérêt liée à une énumération : *veste et cravate*. La réponse attendue était *NIL*, soit celle correspondant à une absence de réponse dans le corpus. Cette bonne absence de réponse a été proposée par 5 systèmes, dont le nôtre mais uniquement dans sa version anglais vers français et cela, seulement à cause d'une erreur de traduction. En effet, notre approche tend à toujours

proposer une réponse (par défaut l'expression apposée la plus longue ou le nom le plus fréquent précédant l'objet de la question). Le seul cas où un *NIL* est retourné survient lorsque l'expression à définir est absente du corpus (comme ce fut le cas à raison pour *Linux* (CLEF06/03), pas encore assez populaire en 1994 et 95, dates des documents de la campagne).

- " Casquette de base-ball , jeans , T-shirt , veste et cravate , tenue de jogging : il est impossible de dresser un portrait-robot de ces candidats à une arme . (LEMONDE95-028295)

5 Perspectives : vers une synthèse de définitions

Un tel processus de réponse à des questions définitoires ouvre une perspective intéressante : il permet d'améliorer la notion de satisfaction en s'essayant à une génération ou plutôt à une synthèse pour la réponse proposée. En effet, les réponses extraites depuis un contexte unique ne sont que très rarement exhaustives (surtout s'il est limité à quelques phrases). Pourtant, et à leur lecture, il apparaît évident qu'il est possible d'améliorer *LA* réponse grâce à l'ensemble des réponses candidates (mais il est alors nécessaire de s'assurer de la qualité de ces réponses candidates, par exemple, depuis des critères fréquentiels ou des coefficients d'association).

Ainsi, lorsque nous cherchons à définir *Airbus* (CLEF06/53), le nom le plus fréquemment associé est *Consortium* et parmi les expressions les plus fréquentes sont *Consortium européen*, *Consortium civil*, et *Consortium aéronautique (Consortium, Programme* et *Consortium européen* ont étés présentées par les systèmes et jugées correctes lors de l'évaluation). D'autre part *Consortium aéronautique européen* apparaît dans le corpus mais épisodiquement. Toutefois, force est de constater que ce dernier semble plus satisfaisant (même si sa fréquence moindre le rend peu sujet à extraction). En outre, il peut être synthétisé depuis les premiers. Il suffit de complémenter le nom le plus fréquemment apposé par tous les adjectifs épithètes qui l'accompagnent dans ses réalisations. Un simple étiquetage morphosyntaxique est suffisant. Enfin, une projection (sur le web ou dans le corpus) des expressions ainsi créées peut permettre de vérifier leur validité du point de vue de leur construction et, notamment, dans ce cas, fixer l'ordre des adjectifs (sauf à définir des règles : une nationalité apparaît souvent en dernier). Il en est de même pour le classique *Bill Clinton* (CLEF06/91) tour à tour *président américain* et *président démocrate* mais pourtant les deux à la fois ou des *navettes Atlantis* (CLEF06/01) et *Challenger* (CLEF06/81), avant tout *spatiales* et *américaines*. Par ailleurs, il peut être opportun de disposer d'une ressource, même limitée, afin, par exemple, de manipuler comme un même concept des noms tels que *Chef* et *Leader* (certaines fonctions reviennent particulièrement dans les QD de ces campagnes), et/ou éviter une éventuelle redite des adjectifs.

Idéalement une telle méthode pourrait être appliquée avec des subordonnées relatives mais ces « *pépites* » d'informations seraient probablement jugées comme inexactes dans le contexte actuel des campagnes, tout en étant susceptibles d'induire plus d'erreurs durant leur génération. Cependant, cela permettrait d'apporter une réponse concise adéquate à la question (fictive) *Qu'est-ce que Bell's Beach ?*, depuis le seul passage du corpus (EQueR) contenant l'expression *Bell's Beach*, puisqu'elle serait alors *(la|une) plage où éclatent les plus grosses vagues d'Australie*. Il est à noter qu'ici, le procédé de réponse implique une mise en relation entre *Beach* et *plage* puis d'utiliser ce qui joue le rôle d'un générique comme antécédent de la proposition relative. En d'autres termes, cela revient à complémenter une sorte de classe/hyperonyme par la proposition relative initialement apposée à l'expression à définir.

C' est du côté de Torquay , petite ville côtière du Victoria au sud -est du pays , située à quelques kilomètres de Bell' s Beach , où éclatent les plus grosses vagues d' Australie , que [...] (LEMONDE99-19935)

6 Conclusion

Dans cet article nous avons étudié l'utilisation de l'apposition pour répondre à des questions définitoires telles qu'elles sont proposées dans les campagnes d'évaluation des systèmes de Questions/Réponses. En effet, dans notre système, les réponses candidates pour ces questions sont extraites depuis leur mise en apposition (ou entre parenthèses) avec les objets à définir. Ensuite, afin de filtrer et retenir la meilleure des expressions apposées comme réponse, un choix est effectué depuis une stratégie impliquant différents indices (principalement fréquentiels) dérivés du voisinage des objets à définir. L'hypothèse forte de notre approche est qu'elle ne nécessite pas de connaissances externes ni même une analyse syntaxique. Évaluée dans le cadre d'une participation à la campagne QA@CLEF-2006, elle trouve environ 80% de bonnes réponses. Cependant une analyse détaillée de quelques cas d'échecs rencontrés nous a amené au constat qu'il n'est pas toujours possible d'aboutir à une réponse. Enfin, et parce que les réponses aux questions définitoires s'y prêtent particulièrement, nous avons esquissé une perspective quant à la synthèse de réponses (depuis les meilleures réponses extraites) pour aller plus avant vers des réponses plus satisfaisantes.

Références

AYACHE C., GRAU B., VILNAT A. (2006). EQueR: The French Evaluation campaign of Question Answering Systems. Actes de *LREC'2006*.

CUI H., KAN M.-Y., CHUA T.-S. (2005). Generic Soft Pattern Models for Definitional Question Answering. Actes de *The 28th ACM SIGIR Conference*, 384-391.

FLEISCHMAN M., HOVY E., ECHIHABI A. (2003). Offline Strategies for Online Question Answering: Answering Questions Before they Are Asked. Actes de *ACL-2003*, 1-7.

GILLARD L., SITBON L., BLAUDEZ E., BELLOT P., EL-BÈZE M. (2006). The LIA at QA@CLEF2006. *The Working Notes for the CLEF 2006 Workshop*.

GREENWOOD M.A., SAGGION H. (2004). A Pattern Based Approach to Answering Factoid, List and Definition Questions. Actes de *The 7th RIAO Conference*, 232-243.

HAN K.-S., SONG Y.-I., RIM H.-C. (2006). Probabilistic model for definitional question answering. Actes de *The 29th ACM SIGIR Conference*, 212-219.

HILDEBRANDT W., KATZ B., LIN J. (2004). Answering Definition Questions Using Multiple Knowledge Sources. Actes de *HLT-NAACL 2004,* 49-56.

MALAISÉ V., DELBECQUE T., ZWEIGENBAUM P. (2005). Recherche en corpus de réponses à des questions définitoires. Actes de la Conférence TALN 2005, 43-52.

PRAGER J., RADEV D., CZUBA K. (2001). Answering What-Is Questions by Virtual Annotation. Actes de *HLT-2001 Conference*, 26-30.

VALLIN A., MAGNINI B., GIAMPICCOLO D., AUNIMO L., AYACHE C., OSENOVA P., PEÑAS A., DE RIJKE M., SACALEANU B., SANTOS D., SUTCLIFFE R. (2006). Overview of the CLEF 2006 Multilingual Question Answering Track. *The Working Notes for the CLEF 2006 Workshop*.

VOORHEES E.M. (2003). Overview of the TREC 2003 Question Answering track, Actes de *The 12th TREC Conference*, 54-68.

VOORHEES E.M. (2005). Chapter 10: Question Answering in TREC. Dans VOORHEES E. M., HARMAN D. (éd.): *TREC Experiment and Evaluation in Information Retrieval.* 233-257.

Caractérisation des discours scientifiques et vulgarisés en français, japonais et russe

Lorraine GOEURIOT[1], Natalia GRABAR[2,3], Béatrice DAILLE[1]
[1] LINA/Nantes
[2] INSERM, UMR_S 872, Eq. 20, F-75006 Paris
Université René Descartes, F-75006 Paris
[3] Health on the Net Foundation, SIM/HUG, Genève, Suisse
natalia.grabar@biomath.jussieu.fr,
{lorraine.goeuriot,beatrice.daille}@univ-nantes.fr

Résumé. L'objectif principal de notre travail consiste à étudier la notion de comparabilité des corpus et nous abordons cette question dans un contexte multilingue en cherchant à distinguer les documents scientifiques et vulgarisés. Nous travaillons séparément sur des corpus composés de documents du domaine médical dans trois langues à forte distance linguistique (le français, le japonais et le russe). Dans notre approche, les documents sont caractérisés dans chaque langue selon leur thématique et une typologie discursive qui se situe à trois niveaux de l'analyse des documents : structurel, modal et lexical. Le typage des documents est implémenté avec deux algorithmes d'apprentissage (SVMlight et C4.5). L'évaluation des résultats montre que la typologie discursive proposée est portable d'une langue à l'autre car elle permet en effet de distinguer les deux discours. Nous constatons néanmoins des performances très variées selon les langues, les algorithmes et les types de caractéristiques discursives.

Abstract. The main objective of our study consists to characterise the comparability of corpora, and we address this issue in the monolingual context through the disctinction of expert and non expert documents. We work separately with corpora composed of medical area documents in three languages which show an important linguistic distance between them (French, Japanese and Russian). In our approach, documents are characterised in each language through their thematic topic and through a discursive typology positioned at three levels of document analysis : structural, modal and lexical. The document typology is implemented with two learning algorithms (SVMlight and C4.5). Evaluation of results shows that the proposed discursive typology can be transposed from one language to another, as it indeed allows to distinguish the two aimed discourses. However, we observe that performances vary a lot according to languages, algorithms and types of discursive characteristics.

Mots-clés : Linguistique des corpus, corpus comparable, algorithmes d'apprentissage, analyse stylistique, degré de comparabilité.

Keywords: Corpus linguistics, comparable corpora, learning algorithms, stylistic analysis, degree of comparability.

1 Introduction

Un corpus comparable est un ensemble de textes qui partagent entre eux un certain nombre de caractéristiques. La première de ces caractéristiques est de rassembler des documents qui portent sur des sujets proches. Par exemple, dans le contexte multilingue, les corpus comparables vont réunir les documents qui ont une thématique commune mais qui ne sont pas des traductions (Bowker & Pearson, 2002). Dans un contexte monolingue, les corpus comparables peuvent contenir les documents qui portent sur le même sujet mais relèvent de discours différents. Le terme *comparable* est donc employé afin d'indiquer que les corpus ont un certain nombre de caractéristiques en commun. Celles-ci peuvent concerner le contexte de création du texte (ex. : la période, l'auteur), mais aussi le texte lui-même (ex. : le thème, le genre). Le choix de ces caractéristiques dépend des objectifs fixés. Il influe sur le *degré de comparabilité* des corpus, notion permettant de quantifier la comparabilité d'un corpus.

Dans notre travail, nous étudions un corpus comparable portant sur le domaine médical dans trois langues à forte distance linguistique : le français, le japonais et le russe. Afin de garantir une meilleure comparabilité des documents de notre corpus, nous prenons en compte une caractéristique classique, relative à la thématique du corpus, mais aussi une caractéristique discursive : distinction entre les documents selon qu'ils relèvent du discours scientifique ou vulgarisé. Nous nous basons ici sur la notion de discours telle qu'elle est définie par (Ducrot & Schaeffer, 1999) : *"Tout ensemble d'énoncés d'un énonciateur caractérisé par une unité globale de thème"*. Pour l'implémentation de cette caractérisation, nous effectuons une analyse stylistique contrastive, inspirée des travaux de (Karlgren, 1998), et proposons une typologie pour la distinction automatique des documents scientifiques et vulgarisés en différentes langues. Cette typologie reprend trois niveaux d'analyse des documents : structurel, modal et lexical. L'objectif principal de notre travail consiste à vérifier si cette typologie, basée sur ces trois niveaux, permet effectivement de caractériser le discours d'un document du Web et d'affiner la notion de comparabilité des documents. En travaillant sur un corpus trilingue, nous pourrons aussi observer le comportement de cette typologie selon les langues. Finalement, cette typologie est utilisée afin d'adapter à notre corpus des algorithmes d'apprentissage qui nous servirons pour classer de nouveaux documents.

Dans la suite de cet article, nous présentons d'abord le corpus étudié (sec. 2), la typologie établie pour la caractérisation des discours scientifique et vulgarisé (sec. 3), et les méthodes utilisées pour la détection automatique de ces discours (sec. 4). Nous présentons ensuite les résultats (sec. 5) et concluons (sec. 6).

2 Collecte et caractéristiques du corpus comparable trilingue

Notre corpus de travail est un corpus comparable comportant des documents en trois langues à grande distance linguistique : le français, le japonais et le russe. Dans le cadre de notre étude, relative à la recherche d'information multi- et translangue, nous souhaitons disposer d'un corpus au degré de comparabilité assez élevé, sans pour autant effacer les diversités culturelles et linguistiques propres à chacune des langues étudiées. Nous situons la comparabilité à deux niveaux :

– le premier niveau avec une thématique commune partagée par les documents en trois langues. Nous avons choisi le domaine médical et, plus précisémment, la thématique "diabète et ali-

mentation" : ce thème touche un large public et présente une garantie potentielle de collecter une diversité de documents sur le web. Par ailleurs, du point de vue applicatif, les corpus comparables trilingues étant en partie dédiés à l'extraction d'informations multilingues, un thème commun renforce la garantie de trouver un vocabulaire et des caractéristiques linguistiques communes dans les langues traitées.
– le second niveau avec le type de discours, scientifique et vulgarisé.

Dans la suite de cette section, nous abordons la méthodologie de constitution du corpus trilingue et donnons ses principales caractéristiques.

2.1 Méthodologie de constitution des corpus

Le corpus de cette étude est un corpus comparable dans les langues française, japonaise et russe. Les documents sont extraits du Web. La démarche de constitution du corpus repose sur trois étapes principales :

1. Recherche de pages web correspondant à la thématique visée ;
2. Sélection des pages pertinentes ;
3. Classement de ces pages selon leur type de discours.

Ainsi, lors de la première étape de recherche des pages web, nous avons utilisé trois approches :

(1) Recherche sur le web à l'aide de moteurs de recherche généraux ;

(2) Recherche interne sur des portails (médicaux) en utilisant le cas échéant les moteurs de recherche propres aux sites ;

(3) Exploitation des liens entre les pages. Les deux premières approches nécessitent l'utilisation de mots clés. Afin d'obtenir un large spectre de documents, les requêtes utilisées sont formées avec des combinaisons variées de mots clés tels que *alimentation*, *diabète* et *obésité* étendus avec leurs synonymes relevés dans les dictionnaires, et les termes équivalents extraits des pages visitées. Notons aussi que dans le cas d'utilisation d'un moteur de recherche spécifique à un portail, les mots clés restent spécifiques à ce portail.

Parmi ces documents, nous avons sélectionné manuellement les documents pertinents pour la thématique visée. Enfin, les pages sélectionnées ont été classées selon le type de discours. Nous utilisons les heuristiques suivantes pour décider du statut scientifique ou vulgarisé d'un document :
– un document scientifique est rédigé par des spécialistes à destination de spécialistes.
– Pour la vulgarisation scientifique, nous distinguons les textes écrits par "le grand public" à destination de tous et les textes écrits par des spécialistes à destination du "grand public". Sans distinguer formellement ces deux degrés de vulgarisation, nous accorderons une plus grande place aux documents écrits par des spécialistes au détriment des discussions sur des forums par exemple. Les documents écrits par les spécialistes s'avèrent en effet être plus riches en vocabulaire et plus complets en contenu.

La classification manuelle est basée sur ces heuristiques et elle est confortée par des éléments supplémentaires comme la nature du site contenant le document ou le vocabulaire utilisé. Notre méthodologie de classification manuelle, qui relève de l'empirique, nous a conduit à ne pas inclure certains documents "ambigus" (documents inclassables ou sur lesquels les avis divergeaient) dans les corpus d'apprentissage.

2.2 Caractéristiques des corpus

Le tableau 1 présente les principales caractéristiques du corpus ainsi constitué : le nombre de documents et le nombre de mots dans chacune des langues et pour chaque type de discours.

	Français		Japonais		Russe	
	SC	VU	SC	VU	SC	VU
Nb. documents	65	183	119	419	45	150
Nb. mots	425 800	267 900			318 596	175126
Nb. caractéres	2 668 783	2 845 114	493 587	1 154 773	2298306	2 165 768

TAB. 1 – Caractéristiques du corpus

Ce corpus rassemble ainsi plus de 1 500 000 mots dans trois langues. Les chiffres donnés pour la langue japonaise correspondent au nombre de caractères. L'ensemble des documents collectésfait appel à plus de 3 alphabets (cyrillique, latin, japonais incluant hiragana, katakana et kanji) et d'encodages différents. C'est pourquoi les textes ont tous été transcodés en Unicode UTF-8, seul codage permettant de traiter les alphabets latin et cyrillique, ainsi que les caractères kanjis japonais. Les documents du corpus appartiennent à différents formats, parmi lesquels on compte les formats usuels du web (html, xhtml, php, etc.), mais aussi d'autres formats (pdf, ps, doc, etc.). Toutes les pages ont été conservées dans leur format original, mais aussi converties en texte brut. Les genres du Web (Bretan *et al.*, 1998) ne sont pas tous représentés dans le corpus français, dans lequel on trouve en majorité des rapports et articles (de presse ou scientifiques), contrairement au corpus japonais dans lequel on trouve une grande diversité (allant du rapport scientifique à l'offre d'emploi). Le corpus russe montre également une variabilité de genres (articles, ouvrages, recettes de cuisine, guides de bonne pratique, discussions sur des forums spécialisés, ...).

3 Une typologie pour l'analyse stylistique du corpus

L'analyse stylistique est une discipline linguistique trouvant son application informatique dans le domaine de la catégorisation textuelle. Deux grands courants se distinguent dans l'analyse stylistique. Les travaux adoptant la *démarche inductive* permettent de faire émerger d'un corpus des corrélations déterminant des classes de similarités (pouvant varier selon le type de caractéristiques utilisé). Cette méthode permet de créer des typologies dites inductives. Dans le cadre de l'apprentissage automatique, cette méthode s'apparente à l'apprentissage non-supervisé, ou clustering (Biber, 1989). La seconde démarche, appelée *démarche déductive*, consiste à analyser un ensemble de documents préclassés afin de caractériser l'appartenance d'un élément à une classe, sous la forme d'une typologie. Cette démarche s'apparente à l'apprentissage supervisé (Bretan *et al.*, 1998). Avec la démarche déductive, deux techniques permettent d'arriver à une typologie : l'analyse des documents un à un, ou l'analyse contrastive de documents appartenant à deux classes distinctes. Dans notre travail, nous avons choisi d'utiliser l'approche contrastive.

Les algorithmes de catégorisation textuelle (voir section 4.1) s'appliquent à de nombreuses typologies. Les plus fréquentes sont les typologies thématiques ou de genres (Bretan *et al.*, 1998). Les travaux portant sur les typologies de discours étant en revanche moins nombreux, nous allons adapter les algorithmes de la démarche déductive et contrastive à la typologie des discours.

Les documents de notre corpus, collectés sur le Web, présentent une structure propre que nous ne pouvons pas négliger. Ainsi, contrairement à un grand nombre de travaux traitant de l'analyse stylistique de textes, par exemple (Malrieu & Rastier, 2002 ; Biber, 1989), nous prenons en compte aussi bien le contenu textuel que la structure des documents. Nous nous basons sur ces deux types d'informations afin de dégager une typologie propre aux discours ciblés.

Sinclair (1996) dans ses travaux typologiques introduit une notion de niveaux dans les typologies textuelles. En effet, il est selon lui plus pertinent de distinguer deux catégories de critères : les critères externes, caractéristiques du contexte de création du texte ; et les critères internes, caractéristiques linguistiques du texte. Notre corpus étant construit à partir de documents issus du Web, nous considérons les critères externes comme étant les critères relatifs à la création du document et à sa structure (caractéristiques "non-linguistiques"). La partie interne concerne les caractéristiques linguistiques du document. Cependant, l'analyse stylistique met en évidence différents niveaux de granularité dans les critères. La distinction entre les documents scientifiques et vulgarisés induit une prise en compte du locuteur dans son discours, c'est-à-dire de la modalité. De plus, le discours scientifique peut se caractériser par le vocabulaire employé, la longueur des mots, et autres critères relevant du lexique. La typologie que nous adoptons distingue donc trois niveaux d'analyse des documents :

Caractéristiques structurelles : éléments de la structure graphique et textuelle du texte ;

Caractéristiques modales : éléments caractérisant la modalité dans le texte ;

Caractéristiques lexicales : éléments relatifs au lexique employé dans le texte.

Critère	Français	Japonais	Russe
Format d'URL	×		
Format de document	×	×	×
Méta-informations (présence)	×	×	×
Titre de la page (présence)	×	×	×
Techniques de mise en page	×	×	×
Fonds	×	×	×
Images	×	×	×
Paragraphes	×	×	×
Listes	×	×	×
Nombre de phrases	×	×	×
Typographie	×	×	×
Longueur du document	×	×	×

TAB. 2 – Caractéristiques structurelles

Les caractéristiques structurelles, présentées dans le tableau 2, concernent en majeure partie l'aspect graphique du document (format, images, fonds, ...) ainsi que les éléments de sa structure pris en compte ici grâce aux balises `HTML` (paragraphes, listes, titre, ...). L'ensemble de critères structurels sont détectables dans les trois langues.

Dans le tableau 3 sont présentées les caractéristiques modales, qui correspondent à la modalité dans les textes, c'est-à-dire à la position du locuteur dans son propre discours. Ces critères sont directement inspirés des théories de Charaudeau (1992), et ont été adapté à notre corpus (Krivine *et al.*, 2006; Nakao, 2006). Parmi les actes locutifs énoncés par Charaudeau (1992), nous n'avons conservé que ceux qui s'avérent opératoires dans les documents analysés. Par exemple, la modalité de l'opinion peut être détectée en français grâce aux verbes comme *penser, paraître, sembler*. Si, dans leur majorité, ces critères se retrouvent dans les trois langues, quelques

	Critère	Français	Japonais	Russe
Modalité allocutive	Pronoms personnels sujets allocutifs	×	×	
	Modalité de l'injonction	×	×	×
	Modalité de l'autorisation	×		×
	Modalité du jugement	×		
	Modalité de la suggestion	×	×	×
	Modalité de l'interrogation	×	×	×
	Modalité de l'interpellation	×		×
	Modalité de la requête	×	×	×
Modalité élocutive	Pronoms personnels sujets élocutifs	×	×	
	Modalité du constat	×	×	×
	Modalité du savoir	×	×	×
	Modalité de l'opinion	×	×	×
	Modalité de la volonté	×	×	×
	Modalité de la promesse	×	×	×
	Modalité de la déclaration		×	×
	Modalité de l'appréciation	×		×
	Modalité de l'obligation	×		×
	Modalité de la possibilité	×		×
	Modalité de l'interdiction			×

TAB. 3 – Caractéristiques modales

Critère	Français	Japonais	Russe
Vocabulaire spécialisé	×	×	×
Caractères numériques	×	×	×
Unités de mesure	×	×	×
Longueur des mots	×		×
Bibliographie	×	×	×
Citations bibliographiques	×	×	×
Ponctuation	×	×	×
Fins de phrases		×	
Parenthèses	×	×	×
Autres alphabets (latin, hiragana, katakana)		×	×
Symboles		×	

TAB. 4 – Caractéristiques lexicales

critères (jugement, interdiction) s'avèrent spécifiques à une ou deux des langues traitées, essentiellement parce que l'instantiation de la typologie était basée sur des études assez isolées des corpus dans chaque langue.

Enfin, dans le tableau 4, nous présentons les critères lexicaux. Plus que dans les deux types précédents, ces critères montrent une dépendance selon les langues, comme l'usage de caractères hiragana ou katakana pour la langue japonaise et celui de l'alphabet latin en russe. Notons aussi que certains de ces critères sont spécifiques des documents scientifiques, comme les bibliographies et citations bibliographiques, le vocabulaire spécialisé ou les unités de mesure.

4 Classification automatique selon les discours

La typologie tripartite (caractéristiques structurelles, modales et lexicales) sert de base aux algorithmes d'apprentissage automatique. Nous présentons dans cette section la méthode : les algorithmes d'apprentissage utilisés et les principes d'évaluation des résultats.

4.1 Apprentissage

Comme nous l'avons annoncé, l'objectif applicatif de notre travail consiste à adapter les algorithmes d'apprentissage à la catégorisation des documents en fonction de deux discours, scientifique et vulgarisé. Dans notre travail, l'analyse stylistique, et la catégorisation des documents, est effectuée à travers la typologie de critères proposée dans la section 3. En effet, les algorithmes d'apprentissage perçoivent les documents sous formes de vecteurs, où chaque élément du vecteur représente la valeur d'un critère pour le document correspondant. La longueur d'un vecteur correspond à la fréquence (brute ou pondérée) de chaque critère dans le document traité. En partant d'un corpus d'apprentissage, où les documents sont répartis en deux échantillons (scientifique et vulgarisé), et d'une liste de critères, les méthodes d'apprentissage génèrent une procédure de classification. Cette procédure est ensuite appliquée à de nouvelles données, pour effectuer de nouvelles classifications.

Il existe différentes techniques de classification automatique de textes (réseaux de neurones, classifieurs de Bayes, séparateurs à vastes marges, etc.) pour lesquelles Sebastiani (2002) a effectué un travail de rassemblement et comparaison. Appliquées sur un corpus de dépêches Reuters, ces techniques ont ainsi montré des performances variables en fonction de l'utilisation des approches supervisée ou non-supervisée, de la taille du corpus, du nombre de catégories, etc. Nous avons choisi d'utiliser les séparateurs à vastes marges SVMlight (Joachims, 2002) ainsi que les arbres de décision C4.5 (Quinlan, 1993). Visant un système de classification rapide, nous avons privilégié, pour l'implémentation des critères, des techniques simples basées sur des patrons lexicaux ou lexicaux-syntaxiques et qui analysent superficiellement les documents et leur contenu.

Le fonctionnement du système de catégorisation repose sur la reconnaissance de formes lexicales dans les documents, que ce soit pour les critères graphiques (reconnus à travers les balises), modaux (reconnus à travers les marqueurs de modalité) ou lexicaux.

4.2 Évaluation

Nous avons utilisé la méthode dite *par validation croisée* (*N-fold cross validation*) (Cornuéjols & Miclet, 2002), qui consiste à diviser le corpus d'apprentissage en n sous-échantillons de tailles égales. On retient ensuite un des n échantillons (celui de numéro i) qui sera utilisé pour la phase de test, les autres servant à l'apprentissage. Une fois les résultats collectés, on réitère cette opération en faisant varier i de 1 à n. En posant $n = 5$, à chaque itération l'échantillon d'apprentissage comporte 80 % de nos documents (en terme de caractères) dans l'échantillon d'apprentissage, les autres 20 % dans l'échantillon de test. Les résultats présentés dans la section 5 sont les moyennes sur les 5 partitionnements.

Par ailleurs, nous utilisons les métriques de précision et de rappel pour évaluer nos résultats :

– le rappel, correspondant au nombre de documents correctement classés dans une classe C sur le nombre de documents appartenant à cette classe ;
– la précision, correspondant au nombre de documents correctement classés dans la classe C sur le nombre de documents classés dans la classe C.

5 Analyse et discussion des résultats

Nous avons donc appliqué les algorithmes *SVMlight* et *C4.5* à notre corpus. Les résultats de la classification pour ces deux algorithmes figurent dans les tableaux 5 et 6.

		Français		Japonais		Russe	
		Préc.	Rapp.	Préc.	Rapp.	Préc.	Rapp.
svml	Scientifique	1,00	0,36	0,20	0,41	1,00	0,52
	Vulgarisé	0,80	1,00	0,72	0,80	0,75	1,00
c4.5	Scientifique	0,89	0,80	0,13	0,12	0,50	0,38
	Vulgarisé	0,91	0,94	0,84	0,86	0,74	0,82

TAB. 5 – Précision et rappel pour chaque catégorie de critères obtenus pour les algorithmes SVM light et C4.5 sur les trois langues

		Français		Japonais		Russe	
		Préc.	Rapp.	Préc.	Rapp.	Préc.	Rapp.
svml	Caractéristiques structurelles	0,90	0,67	0,59	0,71	0,85	0,74
	Caractéristiques modales	0,60	0,50	0,50	0,49	0,28	0,50
	Caractéristiques lexicales	0,91	0,75	0,58	0,53	0,98	0,97
c4.5	Caractéristiques structurelles	0,85	0,85	0,41	0,44	0.62	0,68
	Caractéristiques modales	0,89	0,91	0,39	0,44	0,34	0,68
	Caractéristiques lexicales	0,85	0,85	0,47	0,45	0,45	0,52

TAB. 6 – Résultats pour chaque catégorie de critères

Selon le tableau 5, quels que soient la langue et l'algorithme, les documents du discours vulgarisé sont toujours mieux catégorisés. On remarque ainsi que les résultats obtenus avec les documents en français sont dans l'ensemble assez satisfaisants, avec un rappel moyen de 87%, et une précision moyenne de 90% toutes catégories confondues pour le classifieur C4.5 (soit plus de 200 documents bien classés parmi les 250 du corpus). Les résultats de la classification en japonais sont bons pour les documents du discours vulgarisé, mais assez médiocres pour les documents scientifiques. Finalement, avec les documents russes, on obtient de meilleurs résultats avec le système SVMlight, avec un rappel moyen supérieur à 75 % et une précision de 87%. En ce qui concerne les performances plus faibles de la catégorisation des documents scientifiques en japonais et en russe, cela peut être expliqué par la plus forte proportion de documents vulgarisés pour ces langues, au détriment des documents scientifiques (voir le tableau 1). La normalisation des vecteurs d'apprentissage permet de pallier ces trop grandes variations mais il est évident qu'un corpus d'apprentissage plus grand contient des données plus variées et permet de générer un modèle de la langue plus complet.

Par contre, ce résultat est assez surprenant pour le français : le nombre d'occurrences dans les documents scientifiques y est presque deux fois plus important, même si le nombre de documents scientifiques reste inférieur. Quant aux résultats globalement moins performants en japonais, ils peuvent s'expliquer par une forte disparité dans les genres des documents. Par conséquent, il est plus difficile de caractériser ce type de discours par un ensemble de caractéristiques "stable". Les critères utilisés lors de la classification manuelle (voir sec. 2.1) ne sont peut être pas suffisants dans cette langue et devraient être affinés.

Dans le tableau 6 figurent les résultats obtenus par chacun des deux algorithmes en fonction des catégories de la typologie. Chacune de ces catégories montre son importance dans la typologie. En effet, quel que soit le type de classifieur, les résultats obtenus dans chaque langue indiquent qu'il est possible de classer correctement plus de la moitié de nos documents en tenant compte des critères n'appartenant qu'à une seule catégorie. Cependant, aucune des catégories ne se distingue de façon significative dans ces tests. Par ailleurs, les meilleures catégories ne sont pas les mêmes selon le classifieur utilisé. Ainsi, avec *SVMlight*, ce sont les caractéristiques structurelles et lexicales qui se montrent les plus efficaces dans les trois langues. Par contre, avec *C4.5*, chaque langue privilégie des caractéristiques différentes : la modalité en français, le lexique en japonais et la structure des documents en russe. De manière plus détaillée, nous avons pu également constater que, parmi les critères les plus discriminants, on trouve le type d'URL (format de l'URL : site hospitalier, universitaire, etc.), les pronoms délocutifs et le nombre de phrases narratives pour le corpus français ; la longueur des documents, les pronoms élocutifs et les formules de politesse en fin de phrases pour le japonais ; le titre des documents, les images et le vocabulaire spécialisé pour le corpus russe.

6 Conclusion et perspectives

En partant d'un corpus comparable composé de documents issus du Web en français, japonais et russe, nous avons mené une analyse stylistique et contrastive et avons élaboré une typologie pour la caractérisation des discours scientifique et vulgarisé du domaine médical sur le Web. Cette typologie se base sur trois aspects des documents du Web : l'aspect structurel, l'aspect modal et l'aspect lexical. Notre typologie, implémentée grâce aux algorithmes d'apprentissage des séparateurs à vastes marges (SVMlight) ainsi qu'aux arbres de décision (C4.5) donne des résultats de classification satisfaisants. Nous pouvons ainsi "calculer" le type de discours d'un document du Web. Ces résultats montrent que chacune des catégories de la typologie est discriminante pour le système de catégorisation. En effet, chacune d'elles permet d'obtenir des résultats acceptables, tandis que leur combinaison permet de les améliorer. Le discours d'un document du Web peut donc être caractérisé selon ces trois aspects. Certains des critères de la typologie sont présents quelle que soit la langue, ce qui permet de statuer sur leur caractère "universel". Il semble donc qu'il existe des éléments communs dans les documents en langues différentes qui permettent d'indiquer leur type de discours et que, de manière générale, cette typologie est portable d'une langue à une autre. Une des limites principales, observable dans notre travail, semble provenir de la taille insuffisante des corpus, comme les documents scientifiques en japonais et en russe. Par ailleurs, les faibles résultats de la classification des documents scientifiques japonais sont dus également à la diversité des genres dans ce corpus. Cette constatation nous amène à nous interroger sur la composition et la catégorisation manuelle du corpus japonais. Ainsi, suite à la classification hiérarchique des notions de genres et de discours de (Malrieu & Rastier, 2002), on peut se demander si cette distinction ne permettrait pas d'affi-

ner nos résultats. En respectant la distinction des genres, nous pouvons rendre les catégories d'apprentissage plus homogènes et garantir ainsi un plus fort degré de comparabilité dans les corpus. De plus, suite à notre travail, nous nous interrogeons sur la légitimité de la catégorisation binaire effectuée. Nous pensons qu'il pourrait être intéressant de considérer les catégories scientifiques et vulgarisées des documents comme un continuum. Ceci conduirait à attribuer un "degré de vulgarisation" à chaque document plutôt qu'un type de discours.

Remerciements

Ce travail a été mené dans le cadre du projet DECO, programme CNRS-TCAN 2004-2006 en partenariat avec le NII et l'INaLCO. Nous remercions Estelle Dubreil, Sonia Krivine et Masaru Tomimitsu pour leur participation à la construction du corpus comparable.

Références

BIBER D. (1989). A typology of english texts. *Linguistics*, **27**, 3–43.

BOWKER L. & PEARSON J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. London/New York, Routeledge.

BRETAN I., DEWE J., HALLBERG A., WOLKERT N. & KARLGREN J. (1998). Web-specific genre visualisation. In *Proceedings of the 3rd World Conference on the WWW and Internet*.

CHARAUDEAU P. (1992). *Grammaire du sens et de l'expression*. Hachette.

CORNUÉJOLS A. & MICLET L. (2002). *Apprentissage artificiel : Concepts et algorithmes*. Eyrolles.

DUCROT O. & SCHAEFFER J.-M. (1999). *Nouveau dictionnaire encyclopédique des sciences du langage*. Seuil.

JOACHIMS T. (2002). *Learning to Classify Text using Support Vector Machines*. Kluwer Academic Publishers.

KARLGREN J. (1998). *Natural Language Information Retrieval*, chapter Stylistic Experiments in Information Retrieval. Tomek, Kluwer.

KRIVINE S., TOMIMITSU M., GRABAR N. & SLODZIAN M. (2006). Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais. In *Actes de TALN*, volume vol.1, p. 522–531.

MALRIEU D. & RASTIER F. (2002). Genres et variations morphosyntaxiques. *Traitement Automatique des Langues (TAL)*, **42**(2), 548–577.

NAKAO Y. (2006). étude sémantico-discursive contrastive d'un corpus comparable français-japonais. Master's thesis, Université de Nantes.

QUINLAN J. R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

SINCLAIR J. (1996a). *Preliminary recommendations on Corpus Typology*. Rapport interne, EAGLES (Expert Advisory Group on Language Engineering Standards).

SINCLAIR J. (1996b). *Preliminary recommendations on Text Typology*. Rapport interne, EAGLES (Expert Advisory Group on Language Engineering Standards).

OGMIOS : une plate-forme d'annotation linguistique de collection de documents issus du Web

Thierry HAMON, Julien DERIVIÈRE, Adeline NAZARENKO
LIPN – UMR CNRS 7030
99 av. J.B. Clément, F-93430 Villetaneuse, FRANCE
{Thierry.Hamon,Julien.Derivière,Adeline.Nazarenko}
@lipn.univ-paris13.fr

Résumé. L'un des objectifs du projet ALVIS est d'intégrer des informations linguistiques dans des moteurs de recherche spécialisés. Dans ce contexte, nous avons conçu une plate-forme d'enrichissement linguistique de documents issus du Web, OGMIOS, exploitant des outils de TAL existants. Les documents peuvent être en français ou en anglais. Cette architecture est distribuée, afin de répondre aux contraintes liées aux traitements de gros volumes de textes, et adaptable, pour permettre l'analyse de sous-langages. La plate-forme est développée en Perl et disponible sous forme de modules CPAN. C'est une structure modulaire dans lequel il est possible d'intégrer de nouvelles ressources ou de nouveaux outils de TAL. On peut ainsi définir des configuration différentes pour différents domaines et types de collections. Cette plateforme robuste permet d'analyser en masse des données issus du web qui sont par essence très hétérogènes. Nous avons évalué les performances de la plateforme sur plusieurs collections de documents. En distribuant les traitements sur vingt machines, une collection de 55 329 documents du domaine de la biologie (106 millions de mots) a été annotée en 35 heures tandis qu'une collection de 48 422 dépêches relatives aux moteurs de recherche (14 millions de mots) a été annotée en 3 heures et 15 minutes.

Abstract. In the context of the ALVIS project, which aims at integrating linguistic information in topic-specific search engines, we developed an NLP architecture, OGMIOS, to linguistically annotate large collections of web documents with existing NLP tools. Documents can be written in French or English. The distributed architecture allows us to take into account constraints related to the scalability problem of Natural Language Processing and the domain specific tuning of the linguistic analysis. The platform is developed in Perl and is available as CPAN modules. It is a modularized framework where new resources or NLP tools can be integrated. Then, various configurations are easy to define for various domains and collections. This platform is robust to massively analyse web document collections which are heterogeneous in essence. We carried out experiments on two different collections of web documents on 20 computers. A 55,329 web documents collection dealing with biology (106 millions of words) has been annotated in 35 hours, whereas a 48,422 search engine news collection (14 millions of word) has been annotated in 3 hours and 15 minutes.

Mots-clés : plateforme d'annotation linguistique, passage à l'échelle, robustesse.

Keywords: linguistic annotation, NLP platform, process scability, robustess.

1 Introduction

Si les moteurs de recherche actuels sont suffisants pour répondre aux requêtes les plus courantes sur Internet, il n'existe pas actuellement d'outils permettant la formulation de requêtes s'appuyant sur des techniques de recherche avancées (filtrage sur le sens, élimination d'ambiguïtés, exclusion des sites marchands, etc.) et spécialisées exploitant des connaissances du domaine. Par exemple, la plupart des publications dans le domaine de la biologie et de la bio-médecine sont enregistrées dans de grandes bases de données textuelles, plus ou moins spécialisées (Flybase pour l'espèce Drosophilia Menogaster, Medline pour la biologie et la médecine). Ce type de bases documentaire est aujourd'hui essentiel au travail des scientifiques mais ceux-ci sont confrontés à la masse de textes, sans pouvoir y faire face. Les outils disponibles sont trop généraux, ils renvoient des centaines ou des milliers d'articles pour la moindre requête. Pour juger de la pertinence d'un document dans ce contexte, il faut en analyser le contenu (reconnaissance des entités, reconnaissance des termes techniques).

Le projet ALVIS[1] vise à développer un moteur de recherche *open source* incluant des techniques de recherche avancées et d'analyse du contenu textuel, notamment du point de vue sémantique. Par rapport aux moteurs de recherche actuels, ALVIS cherche à prendre en compte à la fois le thème et le contexte de la recherche pour affiner l'analyse de la requête et du document. Le projet s'appuie sur une architecture *peer-to-peer*. Le système est constitué d'un réseau de « nœuds » assurant l'infrastructure de recherche globale, auxquels sont adjoints des nœuds spécialisés pour un domaine donné. Les nœuds spécialisés proposent une véritable analyse du contenu textuel pour améliorer l'accès au document. A terme, des tâches d'extraction d'informations structurées et leur fusion avec des informations déjà enregistrées au sein de bases de données devraient pouvoir être prises en charge par ce type de moteur spécialisé.

L'accès au contenu sémantique des documents issus du web ou de grandes bases documentaires nécessite une première phase d'enrichissement linguistique des documents en un temps suffisamment court. Il s'agit ici de réduire le goulet d'étranglement que constituent généralement les outils de TAL lorsqu'ils sont intégrés dans des applications de recherche d'information. L'architecture logicielle que nous avons développé permet de satisfaire cette contrainte. Cette plate-forme, OGMIOS, est à la fois générique et spécialisable. Elle est conçue pour analyser de manière robuste des collections de taille variées et hétérogènes du point de vue de la langue (pour l'instant le français et l'anglais[2]), de la longueur et du type de leurs documents . Elle peut aussi être spécialisée pour un domaine particulier. Dans le cadre du projet ALVIS, les expériences ont porté en priorité sur le domaine de la biologie, mais nous avons également pu tester la plate-forme sur un corpus de dépêches relatives aux moteurs de recherche.

Cet article présente notre approche permettant de répondre aux contraintes de performances, de généricité et d'adaptabilité à un domaine de spécialité, qu'impose l'utilisation du TAL dans une application de recherche d'information (RI) spécialisée. Dans la section 2, nous donnons un aperçu de l'état de l'art des plates-formes d'annotation de documents. La plate-forme est décrite dans la section 3 avec les modules de traitement qu'elle intègre. L'évaluation des performances de la plateforme est présentée à la section 4.

[1] ALVIS Superpeer semantic Search Engine, projet IST / STREP n° 002068, voir `http://www.alvis.info/alvis`.

[2] Des versions slovène et chinoise ont également été développées dans le cadre du projet mais avec une ambition moindre pour le slovène et avec une architecture un peu différente pour le chinois.

2 État de l'art

Lors de cette dernière décennie, plusieurs architectures d'ingénierie du texte ont été développées pour articuler les traitements linguistiques (Cunningham *et al.*, 2000) sans toutefois se placer dans un contexte de recherche d'information. Ainsi, les architectures GATE (Bontcheva *et al.*, 2004), UIMA (Ferrucci & Lally, 2004) ou de Textpresso (Müller *et al.*, 2004) visent généralement l'annotation linguistique et l'exploration de corpus de taille moyenne pour l'extraction d'information. LinguaStream (Widlöcher & Bilhaut, 2005), quant à elle, est conçue comme un outil de dépouillement de corpus et d'expérimentation, qui formalise des traitements complexes.

Ces plates-formes appuient leur analyse des documents sur des outils de Traitement Automatique des Langues existants. Ceux-ci sont réutilisés dans des modules qui les encapsulent et qui assurent la conformité des entrées/sorties. La définition d'un format d'échange et d'annotation suffisamment générique est également un point crucial pour les plates-formes d'annotation. Il s'agit d'assurer une communication correcte des informations entre les modules, mais aussi une réutilisation des annotations produites dans des applications externes. Ont ainsi été proposés différents formats d'échange et d'annotation qui reposaient généralement sur SGML puis XML. Le format d'échange et d'annotation de GATE, CPSL (Common Pattern Specific Language) et d'UIMA, CAS (Common Analysis Structure) sont inspirés du format d'annotation TIPSTER (Grishman, 1997). Afin de préserver une certaine flexibilité, les annotations y sont déportées.

Au regard de nos contraintes (généricité, performances et adaptabilité à un domaine de spécialité), les plates-formes d'annotation existantes ne paraissent pas adaptées à la recherche d'information spécialisée. Si les plates-formes GATE et UIMA sont plutôt conçues comme des solutions génériques, le système Textpresso (Müller *et al.*, 2004) poursuit un objectif similaire au nôtre : proposer une architecture générique capable de traiter des corpus de documents issus d'un domaine spécialisé. Cette plate-forme a été conçue pour la fouille des documents traitant de biologie, aussi bien des résumés que des articles complets. Son évaluation a porté sur un corpus relativement petit : 16 000 résumés et 3 000 articles en texte brut.

En règle générale, on dispose de très peu d'informations pour d'évaluer les performances de ces systèmes sur un corpus de documents. Un premier test nous a montré que GATE ne convient pas au traitement de gros corpus de documents : seuls de petits volumes de documents pouvaient être traités sans rencontrer des problèmes. Ceci s'explique par le fait que GATE ait été conçue comme un environnement puissant de développement et de conception d'applications de TAL dans le cadre de l'extraction d'information. Le passage à l'échelle n'était pas un objectif central. La *méta-plate-forme* KIM (Popov *et al.*, 2004), qui s'appuie sur GATE, tente cependant de satisfaire cette contrainte dans le cadre de projets d'annotation sémantique massive SWAN[3] et SEKT[4]. L'architecture est dédiée à l'enrichissement d'ontologies, l'indexation sémantique et la recherche d'information. Bien que les auteurs identifient le passage à l'échelle comme un paramètre critique, aucune performance en terme de temps de calcul et de volume de documents traités, n'est fournie. Le traitement de grande collections de documents est cependant, envisageable avec UIMA, les temps de calcul et l'adaptabilité des traitements restant à évaluer. Celle-ci offre en effet la possibilité de traiter les documents les uns après les autres ou sous forme d'une collection. Le Collection Processing Engine (CPE) gère alors la parallélisation et surveille les performances.

[3] `http ://deri.ie/projects/swan`
[4] `http ://sekt.semanticweb.org`

Les plates-formes existantes répondent donc partiellement aux contraintes de l'intégration d'informations linguistiques dans un moteur de recherche spécialisé. Il s'agit généralement plus d'environnements de dépouillement que d'architectures d'annotation de gros volumes de données pouvant être utilisées pour la recherche d'information spécialisée. Nous avons donc développé une plate-forme capable de gérer d'importants volumes de documents en mettant l'accent sur l'efficacité et la robustesse des traitements effectués.

3 Une plate-forme modulaire et adaptable

Nous avons choisi de développer une plate-forme d'annotation linguistique exploitant des outils de TAL existants plutôt que d'en développer de nouveaux[5]. Nous avons ainsi pu mettre l'accent sur la robustesse des traitements et la rapidité de l'annotation de grandes quantités de documents spécialisés, en proposant une architecture modulaire et distribuée. De plus, l'adaptation des traitements nécessaires à l'analyse de textes spécialisés est réalisée par l'intégration de ressources spécifiques au domaine ou l'utilisation d'outils spécialisés pour un domaine.

3.1 Contraintes spécifiques

Le fait de réutiliser des outils existants et d'autoriser le remplacement de certains outils par d'autres imposent des contraintes spécifiques. Il faut notamment gérer l'hétérogénéité des formats d'entrées/sorties des outils utilisés dans la plate-forme. Chaque outil ayant généralement ses formats propres, il est donc crucial de définir un format d'échange permettant d'interconnecter librement des outils ensemble et de distribuer correctement les traitements.

Le développement d'une plate-forme d'annotation des textes spécialisés intègre également des contraintes spécifiques au TAL, comme la disponibilité de ressources lexicales, terminologiques et ontologiques, ou la nécessité d'adapter des outils au domaine afin d'améliorer certains traitements, comme l'étiquetage morpho-syntaxique ou l'analyse syntaxique, sur des sous-langages particuliers. De plus, toutes les étapes de traitement n'étant pas également pertinentes pour toutes les applications, nous avons préservé au maximum l'approche modulaire.

3.2 Architecture générale

Les différentes étapes de traitement sont traditionnellement prises en charge par un ensemble de modules (Bontcheva *et al.*, 2004). Chaque module est dédié à un type de traitement : reconnaissance d'entités nommées, segmentation en mots, étiquetage morpho-syntaxique, analyse syntaxique, etc. Un module encapsule l'outil effectuant une analyse linguistique donnée et assure la conformité du format des entrées/sorties avec la définition de type de documents (dorénavant DTD). Les annotations sont enregistrées dans un format XML déporté afin de pouvoir mieux gérer l'hétérogénéité des entrées/sorties des outils de TAL. La DTD est décrite dans (Taylor, 2006; Nazarenko *et al.*, 2006). La modularité de l'architecture facilite la substitution d'un outil par un autre, car le remplacement d'un outil n'a aucun impact sur l'ensemble de l'architecture.

[5]Nous avons toutefois développé des outils lorsqu'aucun outil répondant à nos besoin n'était disponible ou nous convenait. Nous avons, de plus, choisi de préférence des logiciels sous licence GPL ou libre/gratuit pour un usage non commercial.

La spécialisation de la plate-forme pour un domaine spécifique est assurée par les ressources de chacun des modules. Par exemple, une liste d'espèces ou de gènes peut être ajoutée au module de repérage d'entités nommées spécifiques à la biologie, afin de traiter des résumés de Medline. L'adaptabilité des traitements peut aussi se faire par l'intégration d'outils spécialisés.

La figure 1 présente l'architecture de la plate-forme dans son état actuel. D'autres modules tels que l'étiquetage sémantique et la résolution d'anaphores seront prochainement intégrés. Les boîtes représentent les différents modules composant la chaîne de traitements linguistiques. Ces modules sont décrits dans la section 3.3. Les flèches en traits pleins représentent le flux de données lors du traitement tandis que les flèches en pointillés représentent les ressources qui peuvent être utilisées dans la plate-forme.

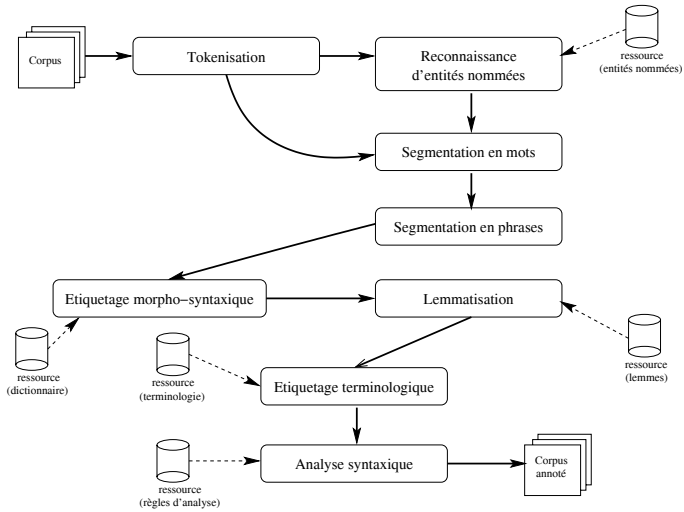


FIG. 1 – Architecture de la chaîne de traitement

Nous partons du principe que les documents Web donnés en entrée ont déjà été téléchargés, nettoyés, codés en UTF-8 et convertis au format XML (Taylor, 2006). Les documents sont d'abord tokenisés, ce qui permet de définir des offsets (indices délimitant une séquence, en nombre de caractères par rapport au début du document) pour garantir l'homogénéité des différentes annotations. Les tokens seront utilisés par les modules suivants. Les documents sont ensuite traités par divers modules : repérage d'entités nommées, segmentation en mots et en phrases, lemmatisation, étiquetage morpho-syntaxique, étiquetage terminologique et analyse syntaxique.

Cette architecture est assez traditionnelle mais certains points méritent d'être soulignés :

– La tokenisation constitue la première étape de la chaîne. Elle procède à une première segmentation, non linguistique, utilisée par la suite par les autres outils. Le token est donc l'unité textuelle de base dans la chaîne de traitements, et n'est qu'un point de départ pour les autres annotations. Ce niveau d'annotation suit les recommandations du groupe TC37SC4/TEI, même si nous employons le terme d'*offset de caractère* plutôt que celui de *pointeur d'élément* pour désigner les frontières de chaque token. Pour simplifier les traitements suivants, nous distinguons quatre types de tokens : alphabétiques, numériques, séparateurs et symboliques.

– L'étiquetage des entités nommées se produit très tôt dans la chaîne de traitement car l'identification des entités nommées facilite la désambiguïsation d'un certain nombre de marques de ponctuation lors de la segmentation en mots ou en phrases.
– L'étiquetage terminologique est utilisé tel quel mais peut également être considéré comme un préalable à l'analyse syntaxique. Cette dernière demandant beaucoup de temps de calcul, nous exploitons le fait qu'une analyse terminologique réduit le nombre d'analyses syntaxiques possibles (Aubin *et al.*, 2005).

3.3 Description des modules disponibles

Les modules sont appelés de manière séquentielle pour chaque document. Les sorties (annotations) sont stockées en mémoire jusqu'à la fin du traitement du document en cours, puis enregistrées dans un format XML.

Cette section décrit les différents modules intégrés à l'heure actuelle au sein de la chaîne de traitement. Il s'agit d'une description des modules par défaut de la plate-forme pour le traitement de l'anglais. Des outils similaires sont également intégrés pour le français (à l'exception de l'analyse syntaxique). De plus, la conception et l'implémentation de la plate-forme permet aisément une substitution d'un outil par un autre.

Etiquetage d'entités nommées. Le module assurant la reconnaissance des entités nommées identifie les séquences textuelles qui renvoient à une entité, leur associe un type sémantique (dépendant du domaine – pour la biologie, les étiquettes `gene` et `species`, par exemple) et, le cas échéant, normalise cette séquence. Dans la suite des traitements, une entité nommée est considérée comme une seule unité et assimilée à un mot. En les reconnaissant à un stade très préliminaire dans l'analyse, on évite des ambiguïtés ultérieures. Le module encapsule TagEN (Berroyer, 2004), qui repose essentiellement sur des dictionnaires et l'application de règles décrites sous formes de transducteurs.

Segmentation en mots et en phrases. Ce module identifie les phrases et les mots. Il exploite un ensemble d'expressions régulières reprenant l'algorithme proposé dans (Grefenstette & Tapanainen, 1994). Une partie de la segmentation est effectuée par le module de reconnaissance des entités nommées dans la mesure où celui-ci résout un grand nombre des problèmes liés à la ponctuation. C'est par exemple le module traitant les entités nommées qui permet de reconnaître la séquence « B. subtilis » , et qui met en rapport l'abréviation « B. » avec la forme étendue « Bacillus » . Le point présent dans la séquence « B. subtilis » n'a plus à être pris en compte au niveau de la segmentation en phrases.

Étiquetage morpho-syntaxique. Ce module associe une étiquette morpho-syntaxique à chaque mot du texte. Il repose sur la segmentation effectuée à l'étape précédente. Nous utilisons à l'heure actuelle le TreeTagger (Schmid, 1997). Nous avons aussi testé l'intégration de l'étiqueteur GeniaTagger (Tsuruoka *et al.*, 2005) qui est spécialisé pour le biologie, même si on observe que le gain en qualité de l'étiquetage se fait au détriment des performances.

Lemmatisation. Ce module associe un lemme à chaque mot du texte. Si le mot ne peut pas être lemmatisé (nombres, mots étrangers, mots inconnus), aucune information n'est associée à la forme. Ce module suppose que l'analyse morpho-syntaxique a préalablement été effectuée. Dans notre implémentation, la lemmatisation est effectuée en même temps que l'analyse morpho-syntaxique par le TreeTagger mais quand on utilise un étiqueteur qui ne fournit pas de lemmes, comme l'analyseur de Brill (Brill, 1995), il faut faire appel à un module spécifique pour la lemmatisation.

Étiquetage terminologique. Ce module vise à repérer les expressions du domaine qui ne sont pas des entités nommées, comme *gene expression* ou *spore coat cell* dans le domaine de la biologie. L'analyse peut être réalisée en projetant les termes fournis en entrée. Ceux-ci peuvent être issus de ressources terminologiques comme Gene Ontology (GO Consortium, 2001), le MeSH (MeSH, 1998) ou UMLS (UMLS, 2003) ou d'une ressource construite à l'aide d'un extracteur de termes. L'analyse morpho-syntaxique et la lemmatisation du texte sont nécessaires pour procéder à l'analyse terminologique.

Analyse syntaxique. L'analyse syntaxique vise à produire, pour chaque phrase du texte, un graphe reflétant les dépendances entre mots au sein de la phrase. L'analyse repose sur les sorties de l'analyse morpho-syntaxique. La plupart des analyseurs n'exigent pas une analyse terminologique préalable mais celle-ci permet de faire décroître largement l'ambiguïté et donc la complexité de l'analyse (Aubin *et al.*, 2005).

L'analyse syntaxique demande encore aujourd'hui des temps de calcul beaucoup plus importants que les autres étapes d'analyse, dans la mesure où elle opère sur un espace de recherche très vaste (tous les mots de la phrase sont potentiellement liés deux à deux). Nous avons choisi d'intégrer le Link Grammar Parser (Sleator & Temperley, 1993), qui repose sur des grammaires de dépendance, comme traitement par défaut. Pour le traitement de textes biomédicaux, l'adaptation de cet outil au domaine de la biologie BIOLG (Pyysalo *et al.*, 2006) est utilisée.

3.4 Implémentation

La plate-forme est implémentée en Perl et est disponible sous forme de modules CPAN (`http://search.cpan.org/~thhamon/Alvis-NLPPlatform-0.3/`). Nous avons utilisé un modèle client/serveur, mais la plateforme peut également traiter séquentiellement et de manière autonome une collection de documents. Dans le contexte d'utilisation client/serveur, chaque client récupère auprès du serveur les documents à traiter les uns après les autres et les analyse. Les documents annotés sont ensuite renvoyés au serveur qui, dans l'ensemble de la chaîne de traitement de recherche d'information d'ALVIS, les envoie au moteur d'indexation.

4 Analyse des performances

La plate-forme que nous avons développée vise à analyser des textes provenant du web pour des moteurs spécialisés dans des domaines techniques. Bien qu'il ne s'agisse pas d'analyse en temps réel, les performances doivent être acceptables. On vise ainsi l'analyse de plusieurs giga-octets de données par jour. Ce type de performances implique une architecture distribuée, qui est par

définition robuste dans la mesure où l'on peut ajouter de nouvelles machines en fonction de la charge. Au-delà des performances, le système doit également être robuste face aux documents fournis en entrée, qui peuvent être très variables quant à leur taille ou leur contenu, notamment quand il s'agit de documents issus du web.

Nous avons mené une expérience d'annotation de deux collections de documents issus du Web. La première collection regroupe 55 329 documents biomédicaux (désormais BIO). La plupart des documents XML ont une taille comprise entre 1 kilo-octet et 100 kilo-octets. La taille du plus grand document est 5,7 méga-octets. La seconde collection comporte 48 422 dépêches relatives aux moteurs de recherche (désormais SEN). La taille des documents varie entre 1 et 150 kilo-octets.

Nous nous sommes placés dans le contexte d'annotation d'un flux de documents venant du Web. Ainsi, nous avons réalisé l'ensemble des traitements jusqu'à l'étiquetage terminologique. Pour l'annotation de la collection BIO, nous avons exploité une liste de 375 000 termes issus du MeSH et de Gene Ontology, Sur la collection SEN, la liste comportait 17 341 termes extraits automatiquement. Nous avons utilisé une liste d'environ 400 000 entités nommées, incluant des noms d'espèce et de gènes sur le corpus BIO, ou des noms de personne, de logiciel et de société sur le corpus SEN.

L'annotation des documents a été distribuée sur vingt ordinateurs. La plupart sont des ordinateurs classiques (de type PC) avec 1 giga-octet de mémoire vive (RAM) et un processeur cadencé à 2,9 ou 3,1 GHz. Nous avons également utilisé un ordinateur avec 8 giga-octets de RAM et deux processeurs Xeon cadencés à 2,8 GHz (processeur Xeon dual core). Le système d'exploitation utilisé est Linux (Debian ou Mandrake). Le serveur et trois clients étaient hébergés sur la machine bi-processeur Xeon. Chaque ordinateur personnel abritait un seul client réalisant l'ensemble de la chaîne de traitement.

Les performances obtenues donnent une bonne idée des performances globales de la plate-forme (une évaluation complète aurait demandé des séries plus importantes de test). Le temps d'exécution de chaque module a été enregistré à l'aide du module Perl `Time::Hires`. Les temps d'analyse sont inscrits dans le fichier XML produit en sortie.

L'annotation de la collection, à l'exception de deux documents, a été effectuée en 35 heures. Le corpus est composé de 106 millions de mots et 4,72 millions de phrases. 147 documents ne contenaient aucun mot, ils n'ont donc pas été analysés au-delà de l'étape de tokenisation. Un des clients a analysé un document composé de 414 995 mots.

Les documents du corpus BIO sont analysés en moyenne en 35 secondes. La génération du fichier XML prend en moyenne 2 secondes supplémentaires. Les étapes les plus coûteuses en temps de traitement sont celles qui demandent le plus de ressources, à savoir la reconnaissance des termes (56 % du temps de traitement global) et la reconnaissance des entités nommées (16 % du total).

Lors ces deux expériences, l'ensemble des documents a été traité sans rencontrer de problème. Les performances obtenues montrent que la plate-forme développée est robuste, et qu'elle peut traiter des grandes masses de textes dans des temps raisonnables. Celles-ci pourraient être encore améliorées par une optimisation du code, et un travail approfondi sur le module d'étiquetage terminologique. Le processus permet une indexation précise de documents spécialisés.

5 Conclusion

Nous avons présenté une plate-forme, OGMIOS, destinée à enrichir des documents issus de domaines spécialisés avec des annotations linguistiques. Les expériences présentées ont porté sur des collections de documents issus du web. Nous avons montré que l'architecture et les modules intégrés à la plate-forme sont adaptés au traitement de textes de langue de spécialité. L'architecture est en outre suffisamment générique pour permettre l'adaptation à d'autres domaines. La plate-forme est actuellement utilisée par d'autres partenaires du projet ALVIS et notamment pour l'annotation de documents issus de bibliothèques numériques en biomédecine.

La stratégie adoptée consiste à réutiliser des modules existants et à les adapter au domaine visé. Ceux-ci peuvent bien évidemment être remplacés par d'autres et les traitements peuvent être enchaînés de différentes façons en fonction du résultat visé. Les modules intégrés sont pour l'instant : la reconnaissance des entités nommées, la segmentation en phrases et en mots, l'analyse morpho-syntaxique et la lemmatisation, la reconnaissance des termes et l'analyse syntaxique. Un module de résolution d'anaphore ainsi que d'autres outils terminologiques seront prochainement intégrés.

Les performances sont le point clé de ce type d'application. Nous avons décrit une implémentation distribuée de la plate-forme permettant le traitement de la collection de documents sur plusieurs machines. Les temps de calcul obtenus sont acceptables pour une tâche de RI.

Remerciements

Ce travail a été réalisé, pour l'essentiel, dans le cadre du projet ALVIS (projet européen IST du 6ème programme cadre – Partenaires : HIIT (Helsinki, Finlande), MIG-INRA (Jouy en Josas, France), LSIR-EPFL (Lausanne, Suisse), ULUND (Lund, Suède), DTU (Copenhague, Danemark), LIPN (Paris, France), JSI (Liubliana, Slovénie), DCSTH (Tsinghua, Chine), IndexData (Copenhague, Danemark), Exalead (France), ALMA Bioinformatica (Madrid, Espagne)). Les données et les exemples fournis ont été obtenus en interaction avec les partenaires du projet. La conception de cette plateforme a bénéficié d'une collaboration de plusieurs années avec le groupe MIG de l'INRA qui a notamment défini le cadre des expériences en biologie.

Références

AUBIN S., NAZARENKO A. & NÉDELLEC C. (2005). Adapting a general parser to a sublanguage. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, p. 89–93, Borovets, Bulgaria.

BERROYER J.-F. (2004). Tagen, un analyseur d"entités nommées : conception, développement et évaluation. Mémoire de D.E.A. d'intelligence artificielle, Université Paris-Nord.

BONTCHEVA K., TABLAN V., MAYNARD D. & CUNNINGHAM H. (2004). Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering*, **10**(3-4), 349–374.

BRILL E. (1995). Transformation-based error-driven learning and natural language professing : A case study in part-of-speech tagging. *Computational Linguistics*, **21**(4), 543–565.

CUNNINGHAM H., BONTCHEVA K., TABLAN V. & WILKS Y. (2000). Software infrastructure for language resources : a taxonomy of previous work and a requirements analysis. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2)*, Athens.

FERRUCCI D. & LALLY A. (2004). UIMA : an architecture approach to unstructured information processing in a corporate research environment. *Natural Language Engineering*, **10**(3-4), 327–348.

GO CONSORTIUM (2001). Creating the Gene Ontology Resource : Design and Implementation. *Genome Res.*, **11**(8), 1425–1433.

GREFENSTETTE G. & TAPANAINEN P. (1994). What is a word, what is a sentence ? problems of tokenization. In *The 3rd International Conference on Computational Lexicography*, p. 79–87, Budapest.

GRISHMAN R. (1997). *Tipster architecture design document version 2.3*. Rapport interne, DARPA.

MESH (1998). Medical subject headings. Library of Medicine, Bethesda, Maryland, WWW page `http ://www.nlm.nih.gov/mesh/meshhome.html`,.

MÜLLER H.-M., KENNY E. E. & STERNBERG P. W. (2004). Textpresso : an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, **2**(11), 1984–1998.

NAZARENKO A., ALPHONSE E., DERIVIÈRE J., HAMON T., VAUVERT G. & WEISSENBACHER D. (2006). The ALVIS format for linguistically annotated documents. In *Proceedings of LREC 2006*.

POPOV B., KIRYAKOV A., OGNYANOFF D., MANOV D. & KIRILOV A. (2004). Kim – a semantic platform for information extraction and retrieval. *Natural Language Engineering*, **10**(3-4), 375–392.

PYYSALO S., SALAKOSKI T., AUBIN S. & NAZARENKO A. (2006). Lexical adaptation of link grammar to the biomedical sublanguage : a comparative evaluation of three approaches. In J. F. SOPHIA ANANIADOU, Ed., *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006)*, p. 60–67, Jena, Germany.

SCHMID H. (1997). Probabilistic part-of-speech tagging using decision trees. In D. JONES & H. SOMERS, Eds., *New Methods in Language Processing Studies in Computational Linguistics*.

SLEATOR D. D. & TEMPERLEY D. (1993). Parsing English with a link grammar. In *Third International Workshop on Parsing Technologies*.

TAYLOR M. (2006). Report on metadata frameworks, including concrete representations, for network nodes and semantic document analyses. ALVIS Deliverable 3.1.

TSURUOKA Y., TATEISHI Y., KIM J.-D., OHTA T., MCNAUGHT J., ANANIADOU S. & TSUJII J. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Proceedings of Advances in Informatics - 10th Panhellenic Conference on Informatics*, LNCS 3746, p. 382–392.

UMLS (2003). UMLS knowledge source. National Library of Medicine.

WIDLÖCHER A. & BILHAUT F. (2005). La plate-forme linguastream : un outil d'exploration linguistique sur corpus. In *Actes de la conférence TALN 2005*, p. 517–522, Dourdan, France.

Les Lexiques-Miroirs. Du dictionnaire bilingue au graphe multilingue

Sébastien HATON, Jean-Marie PIERREL
Laboratoire ATILF Nancy Université – CNRS
44 avenue de la Libération, BP 30687, 54063 Nancy CEDEX
{sebastien.haton,jean-marie.pierrel}@atilf.fr

Résumé. On observe dans les dictionnaires bilingues une forte asymétrie entre les deux parties d'un même dictionnaire et l'existence de traductions et d'informations « cachées », i.e. pas directement visibles à l'entrée du mot à traduire. Nous proposons une méthodologie de récupération des données cachées ainsi que la « symétrisation » du dictionnaire grâce à un traitement automatique. L'étude d'un certain nombre de verbes et de leurs traductions en plusieurs langues a conduit à l'intégration de toutes les données, visibles ou cachées, au sein d'une base de données unique et multilingue. L'exploitation de la base de données a été rendue possible par l'écriture d'un algorithme de création de graphe synonymique qui lie dans un même espace les mots de langues différentes. Le programme qui en découle permettra de générer des dictionnaires paramétrables directement à partir du graphe.

Abstract. Lexical asymmetry and hidden data, i.e. not directly visible into one lexical entry, are phenomena peculiar to most of the bilingual dictionaries. Our purpose is to establish a methodology to highlight both phenomena by extracting hidden data from the dictionary and by re-establishing symmetry between its two parts. So we studied a large number of verbs and integrated them into a unique multilingual database. At last, our database is turned into a "multilexical" graph thanks to an algorithm, which is binding together words from different languages into the same semantic space. This will allow us to generate automatically dictionaries straight from the graph.

Mots-clés : dictionnaires bilingues, traduction automatique, graphe multilingue, algorithme, polysémie verbale, dissymétrie lexicographique.

Keywords: bilingual dictionaries, machine translation, multilingual graph, algorithm, verbal polysemy, lexicographical asymmetry.

1 Introduction

A partir de l'analyse de la polysémie des verbes dans un cadre lexicographique et sur la base des lexiques miroirs dont nous expliquerons précisément le principe, nous avons cherché à établir une méthodologie de récupération de l'intégralité des données de dictionnaires bilingues pour les inclure dans des modèles à des fins d'exploitation informatique. Nous avons travaillé sur trois dictionnaires bilingues, français-anglais, français-espagnol et français-italien et la catégorie verbale a été le matériau central de notre travail. Il est à noter que toutes nos études incluent une composante sur le français et sont destinées à l'informatisation, conformément à la politique scientifique de l'ATILF qui est un laboratoire d'analyse et de traitement informatique de la langue française.

Pour mener à bien cet objectif assez vaste, nous avons procédé en trois étapes que nous décrivons dans cet article. Tout d'abord, nous discuterons de la polysémie dans un cadre lexicographique et nous essaierons de montrer quelles sont les ressources réelles des dictionnaires bilingues ainsi que leurs principales lacunes, pour la plupart inévitables. Dans un deuxième temps, nous présenterons la méthodologie que nous avons développée pour essayer de pallier ces lacunes et pour mettre en valeur les informations cachées ou indirectes que ces ouvrages contiennent. Et enfin, nous présenterons les modèles que nous avons conçus à partir de nos précédentes analyses ainsi que l'algorithme de programmation qui permet de transformer ces modèles en graphes lexicaux exploitables informatiquement.

2 Analyse critique d'une entrée d'un dictionnaire bilingue français-anglais

Le premier travail que nous avons effectué a été de faire une analyse critique et descriptive des dictionnaires bilingues, de leur contenu, de leurs structures et surtout de ce qu'ils recèlent réellement en termes de liens et d'informations cachés. À titre d'exemple, dans l'extrait suivant de l'entrée *to leave* tiré du Robert & Collins français-anglais dernière édition, on peut relever une richesse importante en termes d'informations différents (28 types que nous avons relevés et décrits par ailleurs dans (Haton 2006)) :

leave /li:v/ (vb : prêt, ptp **left**)
VT a *(= go away from) [+ town]* quitter, partir de ; *(permanently)* quitter ; *[+ room, building]* sortir de, quitter ; *[+ person, job, one's husband, wife]* quitter ; *[+ one's children]* abandonner ♦ **he left Paris in 2001** il a quitté Paris en 2001 ♦ **we left Paris at 6 o'clock** nous sommes partis de Paris *or* nous avons quitté Paris à 6 heures…
(…)
e *(Math)* **three from six ~s three** six moins trois égalent trois…
f *(in will) [+ money]* laisser {*to* à} ; *[+ object, property]* laisser, léguer {*to* à}…
VI *(= go away) [person, train, ship etc]* partir, s'en aller ; *(= resign)* partir, démissionner ♦ **to ~ for Paris** *[person, train]* partir pour Paris ; *[ship]* partir *or* appareiller pour Paris…
► **leave off** VI *(*= stop)* s'arrêter ♦ **where did we ~ off?** *(in work, reading)* où nous sommes-nous arrêtés? ♦ **~ off !** arrête !, ça suffit ! *
VT SEP a *(*= stop)* arrêter (*doing sth* de faire qch)…
►**leave out VT SEP** a *(= omit) (accidentally)* oublier, omettre ; *(deliberately)* exclure ; [+ line in text] (also Mus) [+ note] sauter…

La lecture d'un tel dictionnaire ressemble presque à un exercice de spécialistes d'autant qu'il faut se familiariser avec les typographies différentes choisies pour caractériser chacun des types. Ce qui nous intéresse prioritairement dans cet extrait et dans le reste du dictionnaire est donc : quelles sont les informations nécessaires pour caractériser un lien de traduction et comment les modéliser ?

L'autre grand intérêt de cet extrait est qu'il nous apprend beaucoup de choses en plus de nous donner les traductions de *to leave*. Si nous observons à nouveau le début de l'extrait, après les informations phonétiques et grammaticales,

leave /li:v/ (vb : prêt, ptp **left**)
VT a ***(= go away from) [+ town]*** **quitter, partir de ;**

on observe que *quitter et partir de* sont assimilables à des synonymes l'un de l'autre lorsqu'ils traduisent *to leave* avec comme objet direct *town*. De même *to go away from* est présenté comme un synonyme de *to leave* dans les mêmes conditions d'emploi et par voie de conséquence comme une traduction potentielle de *quitter et partir de* et pouvant être traduit par eux. Ainsi, cette simple relation de traduction crée deux liens de synonymie et deux liens de traduction qui sont des informations que nous avons appelées « cachées ou indirectes » dans notre travail, et ce ne sont pas les seules. À la lecture de ce fragment, il est permis de penser que la présence de ces données cachées est un phénomène généralisé et qu'un simple dictionnaire bilingue contient également, dans une certaine mesure, deux dictionnaires de synonymes monolingues. En tout état de cause, il faut que ces liens apparaissent clairement dans notre modélisation.

La richesse cachée des dictionnaires ne s'arrête pas là et nous allons maintenant exposer brièvement le principe des lexiques miroirs qui est une manière de réconcilier les deux parties d'un même dictionnaire bilingue car leur confrontation montre dans les faits que ce sont vraiment deux dictionnaires très différents.

3 Principe des lexiques-miroirs et généralisation de la méthode

Pour bien illustrer le fonctionnement des lexiques miroirs ou inversés (Haton 2003, 2006), nous utiliserons à nouveau l'entrée du verbe *to leave*. Ce que nous appelons le lexique droit (le LD) de *to leave* est l'ensemble des traductions françaises de *to leave* à l'entrée de celui-ci dans la partie anglais-français du dictionnaire et qui contient donc, entre autres, *quitter* et *partir de*. Maintenant, ce que nous appelons le lexique miroir ou inversé (le LI) de *to leave* correspond à l'ensemble des verbes français pour lesquels *to leave* est une des traductions potentielles dans la partie français-anglais du dictionnaire. Nous nous attendions lors de la confrontation des lexiques droit et inversé à observer une certaine symétrie entre les deux. Or, nous avons constaté que les cas de symétrie pour les traductions entre lexèmes représentent nettement moins que 50 % du total. En d'autres termes, lorsqu'un verbe est proposé comme traduction d'un autre dans une partie d'un dictionnaire bilingue, il y a moins d'une chance sur deux que la réciproque soit vraie dans l'autre partie du dictionnaire. Pour effectuer cette comparaison, nous avons construit trois lexiques miroirs complets entre le français et l'espagnol, le français et l'anglais et le français et l'italien, ce qui représente 9000 formes verbales du français et leurs correspondants inversés dans les trois autres langues. Cette dissymétrie est encore plus flagrante si l'on essaie de rattacher les justes acceptions du verbe étudié à ses traductions. Nous prenons comme exemple le verbe *traverser* pour lequel nous

avons extrait du dictionnaire Le Petit Robert toutes les acceptions et nous les avons numérotées :

• 1a- **passer**, **pénétrer de part en part**, **à travers** (*un corps, un milieu interposé)* / 1b- *fig :* **passer**, **pénétrer de part en part**, **à travers** / 2- **se frayer un passage à travers** *(des personnes rassemblées)* / 3- **parcourir d'une extrémité**, **d'un bord à l'autre** (*un espace*) / 4a- **couper**, **aller d'un bord à l'autre** (*une voie de communication*) / 4b- *absolu* : **couper**, **aller d'un bord à l'autre** / 4c- *en parlant de choses mobiles* : **couper**, **aller d'un bord à l'autre** / 5- *choses sans mouvement* : **être**, **s'étendre au travers de** / 6a- **aller d'un bout à l'autre de** *(un espace de temps)* / 6b- **dépasser** *(un état durable)* / 7a- **se présenter à** *(l'esprit)* / 7b- **passer par** *(la tête)* / 8- *équit :* **mettre de travers** / 9- *vx :* **se mettre en travers de**, **s'opposer à**

Ensuite, nous avons fait le relevé des verbes anglais qui entretiennent un lien de traduction avec *traverser* et nous avons associé à chaque lien l'indice correspondant. La deuxième colonne du tableau correspond au LD de *traverser* et il contient le relevé des indices se rapportant aux acceptions de *traverser* si le verbe anglais de la première colonne est proposé comme traduction de *traverser*. À l'inverse, la troisième colonne qui correspond au lexique-miroir de *traverser* contient ces indices lorsque c'est *traverser* qui est la traduction du verbe anglais.

Dico droit	Indice T1	Indice T2	Sens proche	Prolongement du sens
Amble through		2		d'un pas tranquille
Barge through		2		en bousculant
be across	5			
Beetle through		3-4a		en vitesse
belt across		3-4a		à toutes jambes, à toute blinde
Come across		= cross t2	= cross	
Come through	1a			
Cross	3-4a-4c-5-7	3-4		
cross over		indéfini	cross	
cross under	3-4a			par en-dessous
Cut		5	cross, intersect	
cut across	5			
cut right across	5			
Ford	3-4a			à gué
get across		indéfini		*Lit*
get over		indéfini	cross	
get through		indéfini		
go across		indéfini	cross	
go right through	1a			de part en part

go through	1a-6b-6a	6b	suffer, endure	
have across	5			
lie over		5		
live through	6a			
make one's way through	2		passer à travers	
negotiate		indéfini	cross	
occur to	**7**	**7**	**come to mind**	
pass through	6b	1-2-3-4		
run across	5	2-3-4-5		en courant (li)
run right across	5			
Sail		3		en bateau
Shoot through	1a-1b			une douleur / balle (sujet)
Stride across	3	2-3-4		à grands pas
Swim across		3		à la nage
travel down	6a			
travel through		3		
undergo	6b			
Wade across	3-4a	4a		à gué
walk across		indéfini		
zigzag through		indéfini		en zigzaguant

Figure 1 : Tableau de comparaison LD/LI pour la traduction du verbe traverser en anglais

On peut observer dans ces conditions d'analyse qu'il y a un seul cas de symétrie dans ce relevé, celui de la correspondance entre *traverser* et *to occur to,* ce qui représente ici un cas sur quarante-deux. L'étude statistique globale n'a que peu d'intérêt, nous retenons surtout cette très grande asymétrie à l'intérieur d'un même dictionnaire. Même si nous admettons que la symétrie ne peut être la règle, il s'agit tout de même d'un phénomène d'une ampleur très inattendue et il faudra en rendre compte dans les modèles.

Ces deux constats ont eu deux conséquences directes sur notre recherche, l'une entraînant l'autre : tout d'abord, la multiplicité des liens réellement présents dans le dictionnaire ainsi que la difficulté de décrire l'asymétrie au cas par cas par les acceptions font qu'il est quasiment impossible de continuer à avoir une approche sémasiologique du problème, c'est-à-dire de continuer à travailler à partir des lexèmes polysémiques et de leurs dégroupements en acceptions. Il a donc fallu rechercher une démarche plus onomasiologique, c'est-à-dire qui traite d'abord de la signification pour remonter vers le mot. Ensuite, nous nous sommes attachés à rechercher une modélisation du problème dans le but d'automatiser intégralement les liens de traduction directs, cachés et asymétriques avec comme objectif avoué d'autoriser la restitution d'un dictionnaire bilingue entièrement symétrique sans omission d'information. Ce souci d'exhaustivité que nous revendiquons rejoint celui qui sous-tend les travaux de Mel'çuk (1984, 1995) mais sans ajout extérieur pour notre part. Sowa (1992) et Dutoit (2000) adoptent la même démarche dans leurs propres modèles en partant du principe que toute la langue est potentiellement formalisable dans un même espace. Bien entendu, nous voyons cette préoccupation comme un objectif stimulant mais pas comme une fin.

Dans l'optique d'une modélisation efficace, nous avons introduit la notion linguistique de lexie au sens que Pottier puis Mel'çuk et son équipe (1995) lui ont donné, à savoir un mot pris dans une acception bien spécifique. Dans le cadre de notre propre modélisation, une lexie est un mot muni d'un faisceau de traits qui la distinguent de toutes les autres lexies du même mot. Voyons toujours notre exemple de *to leave*, ces traits peuvent être une traduction, un ou plusieurs synonymes liés à la traduction, éventuellement des restrictions sur le choix des collocations sujet et objet, sur les constructions syntaxiques voire sur le registre, le domaine d'emploi et sur certaines informations extra-langagières. Bref, une lexie dans nos modèles est caractérisée par tout ce qui est jugé pertinent pour caractériser une relation de traduction dans un dictionnaire bilingue.

4 Modélisation des données et programmation des graphes

4.1 Modélisation du dictionnaire : le traitement des lexies au cas par cas

Le dernier volet de notre travail concerne la modélisation proprement dite des lexies repérées dans les dictionnaires et – éventuellement – dans les corpus littéraires. La première étape de la modélisation est la transformation en bases de données des dictionnaires bilingues. Revenons à notre fragment d'extrait de *to leave*.

leave /li:v/ (vb : prêt, ptp **left**)
VT a ***(= go away from) [+ town]*** **quitter, partir de ;**

Comme nous l'avons dit, ce fragment contient, outre un lien de traduction, plusieurs liens de synonymie et de traduction présumée qu'il va falloir ordonner. Pour en rendre compte, nous avons effectué une nomenclature de liens de synonymie, lesquels se retrouvent dans l'algorithme final. Trois de ces liens sont directs (traduction directe, synonymie monolingue et traduction réciproque), trois autres sont indirects (synonymie partielle en langue source, partielle en langue cible et traduction indirecte).

Le travail de modélisation d'une lexie réside dans le codage de toutes les informations linguistiques présentes dans la ligne. En l'occurrence, notre fragment induit la définition de plusieurs liens et structures pour les lexies présentes (Haton 2004 & 2006) :

- Deux liens de synonymie de traduction directe (ST) entre *to leave* et *quitter* et entre *to leave* et *partir de*
- Un lien de synonymie partielle en langue source (SPS) entre *to leave* et *to go away from*
- Un lien de synonymie partielle en langue cible (SPTC) entre *quitter* et *partir de*
- Deux liens de synonymie indirecte (SI) entre *to go away from* et *quitter* et entre *to go away from* et *partir de*.
- Une structure de traits pour chacune des quatre lexies *to leave*, *to go away from*, *quitter* et *partir de* dont la plus complète est celle de la lexie source *to leave* qui contient :
- Son lemme : *to leave*
- Sa langue: Anglais
- Sa catégorie grammaticale: Verbe

Ces trois critères sont communs à toutes les lexies du même mot *to leave*.

- Elle contient également la collocation sujet (pas renseignée donc valeur par défaut qui est théoriquement « sans importance d'un point de vue sémantique »), la collocation objet (ici *house*), la construction syntaxique (pas renseignée donc transitif direct), le registre (pas renseigné donc « neutre ») et les informations complémentaires (absentes).
- Trois autres structures sont créées pour les trois autres lexies selon un procédé un peu différent car les informations qui les concernent ne sont pas identiques.

Le tout peut être contenu dans une base de données intermédiaire entre le dictionnaire et le graphe qui peut découler de l'implantation de tous les liens de synonymie entre les lexies. Et bien entendu, pour permettre l'exploitation informatique de tous les objets qui viennent d'être décrits, les liens et les structures sont munis automatiquement d'indices numériques permettant de les réunir ultérieurement par l'intermédiaire du programme de création de graphes. Voilà par l'exemple ce que l'algorithme que nous avons conçu fait sur l'ensemble de notre base de données lexicales en la traitant ligne par ligne. Pour justifier l'existence de plusieurs sources de données distinctes, et notamment la source des corpus, l'algorithme prévoit que lorsque les cases *mot-source* ou *synonymes* ne sont pas renseignées, il ne crée qu'une structure pour la lexie sans lien de synonymie.

Dans leur méthodologie d'élaboration, les graphes que nous nous proposons de générer ressemblent aux graphes construits par l'équipe de Sabine Ploux à l'ISC Lyon (1997, 2003) ainsi qu'au dictionnaire des synonymes du CRISCO à Caen (Manguin & François, 2004).

4.2 Algorithme de transformation en graphe multilingue des lexiques bilingues transformés

Nous présentons ci-après l'algorithme qui a servi pour l'écriture du programme initial. Quelques explications sur la typographie choisie sont nécessaires pour faciliter la lecture de l'algorithme : Les fonctions sont en petites majuscules ; les ensembles d'objets de même type sont notés entre crochets {} ; les ensembles de couples attribut-valeur hétéroclites sont entre parenthèses ( ) ; les opérateurs habituels (boucles, actions, conditions, etc.) sont soulignés ; les constantes en position de valeur sont entre guillemets « » ; Les variables sont écrites avec une majuscule initiale (et peuvent être intégralement écrites en grandes majuscules pour certaines) ; *Nil* représente l'ensemble vide ou l'absence d'élément ; chaque ligne de code active de l'algorithme se termine par un point virgule.
L'algorithme de création de graphes « multilexicaux », car nous n'osons dire pleinement multilingues, a été programmé tel quel il y a quelques mois par Etienne Petitjean à l'ATILF. Nous n'avons évidemment pas eu le temps de tester toutes ses potentialités mais, grâce au test réalisé à partir de l'entrée *abandonner* du Robert & Collins enrichie par l'ensemble de ses traductions et des traductions de ces traductions (4 300 liens de traductions répertoriés) nous savons déjà que la transformation de cette base de données en graphe est acquise (cf. résultats accessibles à l'adresse http://www.atilf.fr/perso/pierrel/telechargement/result.txt). D'autre part, la perspective de disposer de lexiques électroniques correctement balisés nous autorise à espérer l'automatisation complète de toute la chaîne de transformation du dictionnaire en base de données puis de la base de données en graphe, et ce quel que soit le dictionnaire d'origine. La conservation de l'étape intermédiaire « Base de Données » nous semble très importante car c'est le seul espace commun à la fois aux données qui sont automatisables et à celles qui ne le sont pas (notamment les données littéraires originales). La réunion des deux types de données constitue une bibliothèque ou banque lexicale de grande ampleur, ouvrant des perspectives particulièrement riches.

(Dico : Ensemble d'objets) : EXTRACTDONNEES (BD : xls)
(Valeur : Chaîne de caractères) : EXTRACT (TAB[x, y] : tableau)
(Dico : Ensemble d'objets) : AJOUT (Objdico : (Lemme, Cat, Lang, Econstr), Dico : Ensemble d'objets)
(Liens : Ensemble de liens) : AJOUTLIEN (Lien : (Indice, Indice, Type), Liens : Ensemble de liens)
(Econstr : Ensemble de constructions) : AJOUTCONSTR (Eltconstr : (Constr, Collocobj1, Collocobj2, Collocsuj, Registre, Info compl), Econstr : Ensemble de constructions)
Ensemble-d'objets : (Objdico, Liens)
Objdico : (Lemme, Cat, Lang, Econstr)
Liens : {lien}
Lien : (Indice, Indice, Type)
(Econstr : ensemble de constructions) : {Eltconstr}
Eltconstr : (Constr, Collocobj1, Collocobj2, Collocsuj, Registre, Info compl)

Dico = {Nil} ; Liens = {Nil} ;
Ligne = 1 ;
Motsource = TAB[ligne, 1] ;
LS = TAB[ligne, 2] ;
Numcons = 0 ;
Tant que1** motsource .NE. Nil **répéter
    Econstr = {Nil} ;
    Tant que2** (Motsource, LS)= (TAB[ligne, 1], TAB[ligne, 2]) **répéter
        Numcons = Numcons + 1 ;
        IndiceMS = Numcons ;
        EltconstrMS = (Constr = EXTRACT (TAB[ligne,10]), Collocobjet1 = EXTRACT (TAB [ligne, 7]), Collocobjet2 = EXTRACT (TAB [ligne, 8]) , Collocsuj = EXTRACT (TAB [ligne, 9]), Registre = EXTRACT (TAB [ligne, 12]), Infoscompls = EXTRACT (TAB [ligne, 13]), IndiceMS) ;
        MC = EXTRACT (TAB[ligne, 3]) ;
        LC = EXTRACT (TAB[ligne, 4]) ;
        SMC = EXTRACT (TAB[ligne, 6]) ;
        SMS = EXTRACT (TAB[ligne, 5]) ;
        IndiceMC = Numcons ;
        Numcons = Numcons + 1 ;
        EltconstrMC = (Constr = EXTRACT (TAB[ligne,11]), Collocobjet1 = EXTRACT (TAB [ligne, 7]), Collocobjet2 = EXTRACT (TAB [ligne, 8]) , Collocsuj = EXTRACT (TAB [ligne, 9]), Registre = EXTRACT (TAB [ligne, 12]), Infoscompls = EXTRACT (TAB [ligne, 13]), IndiceMC) ;
        EconstrMC = {EltconstrMC} ;
        AJOUTLIEN (Liens, (IndiceMS, IndiceMC, « ST ») ;
        Objdico = (Lemme = MC ; Cat = « V » ; Lang = LC ; Constr = EconstrMC) ;
        Dico = AJOUT (Dico, Objdico) ;

        Si** SMS .NE. Nil **alors début
            IndiceSMS = Numcons ;
            Numcons = Numcons + 1 ;
            EltconstrSMS = (IndiceSMS) ;
            EconstrSMS = {EltconstrSMS} ;

AJOUTLIEN (Liens, (IndiceSMS, IndiceMC, « SI ») ;
AJOUTLIEN (Liens, (IndiceSMS, IndiceMS, « SPS ») ;
objdico = (lemme = MC ; cat = « V » ; lang = LC ;
Constr = EconstrMC) ;
Dico = AJOUT (Dico, objdico) ;
fin

Si *SMC .NE. Nil* ***alors début***
IndiceSMC = Numcons ;
Numcons = Numcons + 1 ;
EltonstrSMC = (IndiceSMC) ;
EconstrSMC = {EltconstrSMC] ;
AJOUTLIEN (Liens, (IndiceMS, IndiceSMC, « ST ») ;
AJOUTLIEN (Liens, (IndiceSMC, IndiceMC, « SPTC ») ;
AJOUTLIEN (Liens, (IndiceMC, IndiceSMC, « SPTC ») ;
Objdico = (Lemme = SMC ; Cat = « V » ; lang = LC ;
Constr = EconstrSMC) ;
Dico = Ajout (Dico, Objdico) ;
fin

Si *SMS .NE. Nil* ***et*** *SMC .NE. Nil* ***alors*** *AJOUTLIEN (Liens, (IndiceSMS, IndiceSMC, « SI »)*
EconstrMS = AJOUTCONSTR(EconstrMS, EltconstrMS)
ligne = ligne +1
Fintantque2
objdico = (lemme = Motsource ; cat = « V » ; lang = LS ; Constr = Econstr)
Dico = Ajout (Dico, objdico)
Fintantque1

5 Conclusion

Bien sûr, si notre algorithme est capable de transformer un dictionnaire en graphe, l'inverse est également vrai moyennant quelques aménagements mineurs. De ce point de vue, nous pensons que notre méthode permettra la génération automatique de dictionnaires selon un paramétrage très souple. Par exemple, admettons que nous avons saisi l'intégralité de notre dictionnaire bilingue et que nous voulons le restituer sous une forme parfaitement symétrique. Deux lignes de code informatique suffisent pour demander à ce que tous les liens de synonymie de traduction, naturellement orientés, soient désorientés et le dictionnaire bilingue symétrique est prêt à être généré.

Les applications les plus directes que nous envisageons sont de deux types : nous espérons pouvoir générer au moins partiellement des dictionnaires inédits à partir de ceux qui auront déjà été transformés. Par exemple, admettons que nous ayons saisi un dictionnaire français-anglais et un dictionnaire français-italien. À partir de cela, nous souhaitons réfléchir aux moyens informatiques à mettre en œuvre pour générer un dictionnaire anglais-italien complètement inédit (cf. Teeraparbseree, 2004). Ce travail a déjà été entamé et les premiers résultats sont encourageants, nonobstant la surgénération de liens de synonymie entre l'anglais et l'italien via le français. D'autre part, la méthode semble offrir un certain intérêt pour l'aide à l'alignement des textes en plusieurs langues, ce dernier générant un feedback très profitable pour l'enrichissement du graphe multilingue. En définitive, le dictionnaire

« fusionné » et les textes alignés constituent des mémoires de traduction quasiment inépuisables et susceptibles de s'enrichir mutuellement.

Références

DUTOIT D., (2000). *Quelques opérations Sens → Texte et Texte → Sens utilisant une Sémantique Linguistique Universaliste a priori.* Thèse soutenue à l'Université de Caen, spécialité Informatique Linguistique.

HATON S., (2003). « Les Champs Sémantiques Multilingues Unifiés ». Actes du congrès international *Représentation du Sens*, Montréal.

HATON S., (2004). « Sens, polysémie et multilinguisme : comment générer des champs synonymiques à partir de dictionnaires de langues ». Actes du colloque JETOU2003 *autour du sens*, Toulouse.

HATON S., (2006). *Analyse et modélisation de la polysémie verbale dans une perspective multilingue : le dictionnaire bilingue vu dans un miroir*. Thèse de doctorat soutenue le 25 novembre 2006. Université de Nancy 2.

MANGUIN J-L., FRANÇOIS J., ALII, (2004). *Le dictionnaire électronique des synonymes du CRISCO. Un mode d'emploi à trois niveaux.* Cahiers du CRISCO n°17, juillet 2004. Université de Caen.

MEL'ÇUK I., ALII, (1984). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches Lexico-Sémantiques I.* Les Presses de l'Université de Montréal.

MEL'ÇUK I., CLAS A., POLGUÈRE A., (1995). *Introduction à la lexicologie explicative et combinatoire.* Champs linguistiques. Éditions Duculot, AUPELF UREF.

PLOUX S., (1997). « Modélisation et traitement informatique de la synonymie ». *Linguisticae Investigationes, Tome XXI/1997, Fascicule 1.*

PLOUX S., JI H., (2003). "A Model for Matching Semantic Maps Between Languages (French / English, English / French)". *Computational Linguistics*. 29(2), pp.155-178.

SOWA J., (1992). "Logical structures in the Lexicon" in *Lexical Semantics and Commonsense Reasoning*, edited by James Pustejovsky and Sabine Bergler, LNAI 627, Springer-Verlag, Berlin, 1992, pp.39-60.

TEERAPARBSEREE A., (2004). « Un système adaptable pour l'initialisation automatique d'une base lexicale interlingue par acceptions ». RECITAL 2004, Fès, 21 avril 2004.

Traduction, restructurations syntaxiques et grammaires de correspondance

Sylvain KAHANE
Modyco, Université Paris 10 - Nanterre & CNRS
sk@ccr.jussieu.fr
http://www.kahane.fr

Résumé. Cet article présente une nouvelle formalisation du modèle de traduction par transfert de la Théorie Sens-Texte. Notre modélisation utilise les grammaires de correspondance polarisées et fait une stricte séparation entre les modèles monolingues, un lexique bilingue minimal et des règles de restructuration universelles, directement associées aux fonctions lexicales syntaxiques.

Abstract. This paper presents a new formalisation of transfer-based translation model of the Meaning-Text Theory. Our modelling is based on polarized correspondence grammars and observes a strict separation between monolingual models, the bilingual lexicon and universal restructuring rules, directly associated with syntactic lexical functions.

Mots-clés : traduction automatique, paraphrase, restructuration syntaxique, TST (Théorie Sens-Texte), grammaire de dépendance, fonction lexicale, lexique bilingue, GUP (Grammaire d'Unification Polarisée), grammaire de correspondance, grammaires synchrones.

Keywords: machine translation, paraphrase, syntactic restructuring, MTT (Meaning-Text Theory), dependency grammar, lexical function, bilingual lexicon, PUG (Polarized Unification Grammar), correspondence grammar, synchronous grammars.

1 Introduction

Cet article s'intéresse à la modélisation des restructurations syntaxiques dans un système de traduction par transfert. L'architecture que nous considérons a été introduite dès les premiers travaux en traduction automatique [TA] (Kulagina & Mel'čuk 1967) et reprise par différents systèmes comme la plateforme Etap-2 (Apresian *et al.* 2003), le système TransLex (Nasr *et al.* 1997, Lavoie *et al.* 2000) ou le projet Eurotra (Arnold & des Tombes 1987, Danlos & Samvelian 1992). Cette architecture se caractérise par une réduction au minimum du lexique bilingue ; les règles qui permettent les restructurations sont universelles et séparées des grammaires monolingues.

Notre contribution se situe uniquement à un niveau théorique et formel. Les exemples que nous étudions sont très bien connus (Tesnière 1959 (chapitres sur la métataxe), Lindop & Tsujii 1993, Dorr 1994) et notre travail n'a fait l'objet d'aucune implémentation. Il nous

semble néanmoins que même si la plupart de ces phénomènes sont considérés dans de nombreuses études, il est encore possible d'améliorer la modélisation des traductions mettant en jeu de telles restructurations syntaxiques. Bien que de nombreux systèmes à transfert aient été développés, peu de travaux proposent un formalisme propre pour l'écriture du module de transfert. La plupart des systèmes utilisent processus de transfert ad hoc basé sur de la réécriture d'arbre et des transformations d'arbres. Les formalismes les plus achevés que je connaisse utilisent des grammaires synchrones (Abeillé *et al.* 1990, Nesson *et al.* 2006 ou Ding & Palmer 2005), mais de telles architectures n'opèrent pas une séparation forte entre lexique bilingue et grammaires monolingues, puisque ce sont les régles des grammaires monolingues qui sont alignées.

Notre modélisation s'inscrit dans le cadre de la Théorie Sens-Texte [dorénavant TST] (Žolkovskij & Mel'čuk 1967, Mel'čuk 1988a), laquelle théorie linguistique a d'ailleurs été initialement élaborée dans la perspective de TA. La modélisation des restructurations en TST a été traitée en détail dans Mel'čuk 1988b et plus récemment dans Mel'čuk & Wanner à paraître. La formalisation que nous proposons, utilisant les grammaires de correspondance polarisées développées dans Kahane 2004 et Kahane & Lareau 2005 nous semble à la fois plus simple, plus rigoureuse et plus aboutie.

Nous commencerons par présenter, dans la Section 2, la notion centrale d'unité significative, une nouvelle définition de la structure syntaxique profonde et le principe de la traduction par commutation d'unités significatives. La Section 3 sera consacrée aux restructurations syntaxiques et au lexique bilingue dont on a besoin en conséquence. Les règles de restructuration et la grammaire de correspondance polarisée qui les met en jeu seront introduites à la Section 4. La Section 5 présentera le problème des restructurations multiples et l'algèbre des fonctions lexicales syntaxiques. Nous finirons avec le problème classique des verbes de mouvement à la Section 6.

2 Traduction, unités significatives, structure syntaxique profonde

Traduire, c'est exprimer à peu près le même sens dans une autre langue. Dans la quasi-totalité des cas, il est possible d'obtenir une traduction convenable en remplaçant chaque unité significative de l'énoncé dans la langue cible par une unité significative de la langue source exprimant un sens similaire (et a priori rarement identique, d'où l'impossibilité de trouver une interlingua satisfaisante, même pour un unique couple de langues). En fait, les seuls cas où ceci n'est pas possible sont les cas où l'une des unités significatives de l'énoncé source ne peut pas être traduit de façon satisfaisante, cas qui échappent pour l'instant à toute tentative de traduction automatique.

Les *unités significatives* [US] (Martinet 1960 : chap. 4 ; Ducrot 1995) sont les signes linguistiques indécomposables dans leur signifié[1] : elles se répartissent en *unités lexicales* (y compris les locutions), *unités grammaticales* et *constructions* (au sens des grammaires de construction ; Goldberg 1995). Les US sont les unités de choix ; par exemple dans l'énoncé *La moutarde me monte au nez*, au sens de 'Je sens la colère m'envahir', il y a 4 choix faits par le locuteur et autant d'unités significatives : la locution LA MOUTARDE MONTER AU NEZ , le pronom MOI (sous sa forme atone *me*) qui est l'unique actant de cette locution, le temps

[1] Les signes linguistiques indécomposables dans leur signifiant sont les morphes.

présent et la construction déclarative (nous ne considérerons pas cette US dans la suite, mais elle est bien réelle et s'oppose par exemple à la construction interrogative).

Bien que la *structure syntaxique profonde* [SSyntP] soit présentée dans de nombreuses études en TST (Mel'čuk 1988a), elle nous semble devoir faire l'objet d'une meilleure définition et de quelques aménagements. La SSyntP d'un énoncé est la structure qui indique comment les US de cet énoncé se combinent. Elle peut être vue comme la structure de dérivation de l'interface syntaxe-sémantique (Kahane 2003). La syntaxe de surface [SyntS] impose une structure hiérarchique, squelette sur lequel nous indiquerons les numéros différenciant les différents actants. Relations syntaxiques de surface, régimes et accords n'apparaissent pas dans la SSyntP et peuvent être recalculés à partir du lexique et de la grammaire. Les unités lexicales [UL] (que nous notons en petites capitales, en les faisant suivre de leur partie du discours [pdd]) occupent les nœuds de l'arbre SyntP. Les unités grammaticales [UG] sont indiquées en indice de l'UL avec laquelle elles se combinent. Une des difficultés pour la SSyntP est que les US peuvent se combiner différemment au niveau sémantique [Sém] et syntaxique. Or la SSyntP, qui sert justement d'interface entre Sém et SyntS, se doit d'encoder simultanément les deux. Prenons l'exemple d'un verbe à montée comme SEMBLER dans *Max semble dormir*. Au niveau sémantique, en raison de la synonymie avec *Il semble que Max dorme*, on peut affirmer que 'sembler' n'a qu'un argument qui est 'dormir', qui a lui-même pour argument 'Max'. Par contre, au niveau syntaxique, MAX est le sujet du verbe SEMBLER et pas celui de DORMIR (d'où la métaphore de la « montée » du sujet potentiel de DORMIR en sujet de SEMBLER). Dans la SSyntP, nous indiquerons donc que SEMBLER a deux actants, que nous numérotons 1 et 2, mais que le 1er actant est « additionnel », ce que nous notons par un +, c'est-à-dire qu'il n'apparaît qu'en SyntS. Par ailleurs, nous indiquons, par un exposant sur la pdd (V^1), que le 1er argument de DORMIR n'est pas réalisé comme actant de DORMIR et, par un 1 en exposant sur la dépendance entre son gouverneur et lui (2^1), qu'il est réalisé comme 1er actant de son gouverneur.

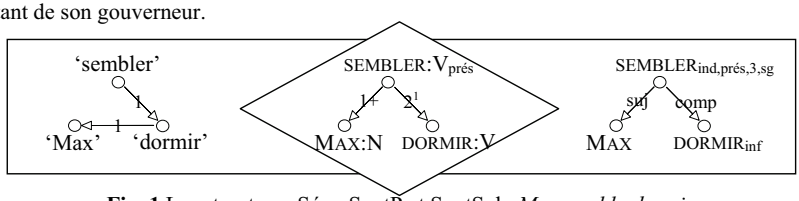
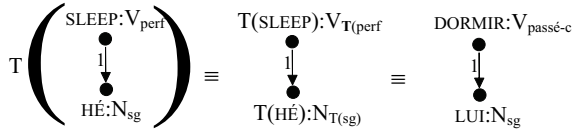


Fig. 1 Les structures Sém, SyntP et SyntS de *Max semble dormir*

En remplaçant dans un énoncé, une US par une US équivalente (c'est-à-dire ayant à peu près le même sens), on obtient une paraphrase. Une traduction sera obtenue en commutant simultanément toutes les US d'un énoncé de la langue source par des US de la langue cible équivalentes[2]. Lorsque chaque US de l'énoncé source est traduite par une US de construction équivalente, la traduction est particulièrement simple : les deux énoncés ayant exactement la même SSyntP. Nous illustrons par la traduction anglais-français suivante :

[2] Nous laissons de côté la question des collocatifs, par ailleurs fort bien traitée dans le cadre de TST. Un *collocatif* est une UL dont le choix est lexicalement contraint par une autre UL, appelée la *base* de la *collocation*. Par exemple pour traduire *gros fumeur*, il faut savoir que GROS est un collocatif de FUMEUR. Le lexique du français devra donc indiquer que Magn(FUMEUR) = GROS, où Magn est la fonction lexicale qui associe à une base un intensifieur. Pour une traduction en anglais, on traduira FUMEUR par SMOKER et on consultera le lexique de l'anglais pour connaître Magn(SMOKER) : T(*gros fumeur*) = T(FUMEUR □ Magn) = T(FUMEUR) □ T(Magn) = SMOKER □ Magn = *heavy smoker*.

(1) **a.** *He slept*.
b. *Il a dormi*.



Cette traduction est réalisée en deux étapes. La première étape modélise notre postulat : la traduction d'un énoncé peut être réalisé en traduisant chaque US séparément. La deuxième étape commute chaque US avec une de ses traductions possibles en utilisant le lexique bilingue anglais-français :

$L_{E\text{-}F}$ =	(SLEEP:V, DORMIR:V)
	(HE:N, LUI:N)
	($:V_{perf}$, $:V_{passé\text{-}c}$)
	($:N_{sg}$, $:N_{sg}$)

$T_{F\text{-}E}$(SLEEP) = DORMIR
$T_{F\text{-}E}$(HE) = LUI
$T_{F\text{-}E}$(perf) = passé-c
$T_{F\text{-}E}$(sg) = sg

Le lexique bilingue (colonne de gauche) a pour ces cas « simples » la forme d'une liste de couples d'unités significatives qui sont la traduction l'une de l'autre. Autrement dit, dire que (SLEEP:V, DORMIR:V) est dans le dictionnaire English-Français revient à dire que $T_{E\text{-}F}$(SLEEP) = DORMIR ou que $T_{F\text{-}E}$(SLEEP) = DORMIR, où $T_{E\text{-}F}$ et $T_{F\text{-}E}$ sont des fonctions de traduction[3].

3 Restructurations et lexique bilingue

Considérons les exemples classiques suivants :

(2) **a.** *I miss you*.
b. *Tu me manques*.

(3) **a.** *Ich schwimme gern*. (litt. je nage volontiers)
b. *J'aime nager*.

(4) **a.** *Zoe needs a book*.
b. *Zoé a besoin d'un livre*.

Pour la traduction (2), on doit non seulement indiquer que MISS et MANQUER sont la traduction l'un de l'autre, mais aussi qu'une conversion des actants est opérée. Pour (3), le lexique allemand-français devra indiquer que GERN est un adverbe[4] dont le gouverneur verbal correspond au deuxième actant de AIMER, tandis que le sujet de ce gouverneur verbal correspond au sujet de AIMER. Nous pouvons modéliser cela de deux façons : soit en indiquant les diathèses dans le lexique bilingue (à la Nasr *et al.* 1997), soit en exprimant ce changement de diathèse par une fonction lexicale (à la Mel'čuk & Wanner 2006). Pour la diathèse, nous utilisons les notations proposées par Kahane & Polguère 2001 pour l'encodage explicite des fonctions lexicales : [*x,y*] désigne la liste des actants par ordre de saillance croissante (*x* est donc le premier actant ; pour un verbe c'est le sujet). Pour un Adj ou un Adv,

[3] $T_{E\text{-}F}$ et $T_{F\text{-}E}$ ne sont pas réellement des fonctions, car la correspondance bilingue est en fait multivoque. Par ailleurs, les dictionnaires bilingues ne sont généralement pas considérés comme symétriques. Nous pensons pour notre part qu'ils doivent l'être et que le fait qu'une traduction dans un sens soit plus utile que dans l'autre provient d'informations monolingues sur l'usage des unités significatives en correspondance.

[4] Plus précisément, GERN est ce que nous appelons un Adv*, c'est-à-dire un Adv qui contrôle le sujet de son gouverneur et a donc deux arguments sémantiques.

le premier « actant » est en fait le gouverneur syntaxique, ce que nous signalons par le symbole ^ (à lire comme une flèche vers le haut).

(MISS:V[1,2], MANQUER:V[2,1]) (GERN:Adv•[2[1]^], AIMER:V[1,2]) (NEED:V, BESOIN:N)	$T_{E\text{-}F}$(MISS) = $Conv_{12}$(MANQUER) $T_{D\text{-}F}$(GERN) = $Adv^1{}_2$(AIMER) $T_{E\text{-}F}$(NEED) = V_0(BESOIN)

$Conv_{12}$ effectue la conversion des actants 1 et 2. $Adv^1{}_2$ adverbialise son mot-clé : le 2 en indice indique que cet Adv va modifier le 2ème actant du mot-clé, le 1 en exposant indique qu'il contrôle aussi le 1er actant de son gouverneur[5]. V_0 est une verbalisation sans changement de diathèse. Pour (4), le lexique bilingue indique que la traduction du verbe NEED n'est pas un verbe mais le nom BESOIN et c'est tout ; c'est le lexique du français qui indique par ailleurs que ce nom possède le verbe support : $Oper_1$(BESOIN:N) = AVOIR.

Voyons maintenant comment s'effectuent les restructurations nécessaires pour ces trois exemples. A chaque fonction lexicale syntaxique est associée une *règle de restructuration* (Mel'čuk 1988). Nous utilisons pour écrire ces règles le formalisme des Grammaires d'Unification Polarisées [GUP] (Kahane 2004) et plus précisément les grammaires de correspondance polarisées (Kahane & Lareau 2005). Trois types d'objets[6] sont considérés : les nœuds syntaxiques représentés par des ronds, les dépendances syntaxiques entre nœuds représentées par des flèches et les liens de correspondance entre nœuds, représentés par des traits pointillés assortis d'un losange pour la polarisation. Chaque objet reçoit une polarité sous la forme d'une couleur noire ou blanche : les objets noirs sont ceux qui sont manipulés par la règle, tandis que les objets blancs expriment des besoins et doivent obligatoirement être fusionnés avec un objet noir. La correspondance est achevée lorsque, après fusion, tous les objets sont noirs.

Les règles associées aux fonctions lexicales $Conv_{12}$ et $Adv^1{}_2$ manipulent uniquement des dépendances[7] : les nœuds syntaxiques, ainsi que les liens de correspondance entre eux sont donc blancs et doivent être construits par d'autres règles. La règle pour $Oper_1$ introduit le nœud supplémentaire pour le verbe support $Oper_1$ et construit un lien[8] (Fig. 2).

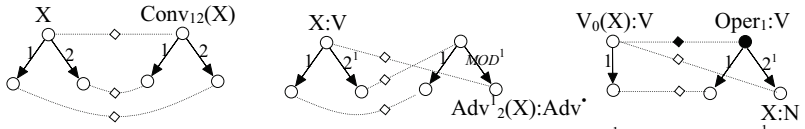


Fig. 2. Règles de restructuration associées à $Conv_{12}$, $Adv^1{}_2$ et $Oper_1$

---

[5] Donnons un autre exemple monolingue : $Adv^1{}_2$(PRÉCIPITER) = PRÉCIPITAMMENT (*Max a précipité son départ* ≡ *Max est parti précipitamment*).

[6] Par souci de simplification, nous n'avons pas représenté les grammèmes comme des objets, mais il le faudrait. Par ailleurs, bien qu'elle l'assure en partie, la grammaire de correspondance n'a pas pour objet de vérifier la bonne formation des SSyntP qu'elle met en correspondance. Ceci peut et doit être assuré par une grammaire indépendante. Voir Kahane & Lareau 2005 pour l'interfaçage de grammaires de correspondance et de grammaires de bonne formation.

[7] Nous notons *MOD* la relation modificative (plutôt que *ATTR* comme il est usuel en TST). Il s'agit généralement d'une relation prédicative dont le prédicat est le dépendant. Le 1 en exposant indique que le 1er actant du gouverneur est également un argument du dépendant.

[8] Le fait que V_0(X) soit liés à la fois à X et à $Oper_1$ est utile. En effet, un modifieur peut aussi bien aller sur X (*Max a* ***grand*** *besoin d'un livre*) que sur $Oper_1$ (*Max a* ***vraiment*** *besoin d'un livre*).

La correspondance entre deux SSyntP s'effectue en combinant les règles de restructuration et des *règles de copie* assurant la copie des unités lexicales, unités grammaticales et dépendances qui n'ont pas encore été prise en compte (Fig. 3)

Fig. 3. Règles de copie pour AIMER, le présent et une dépendance 1

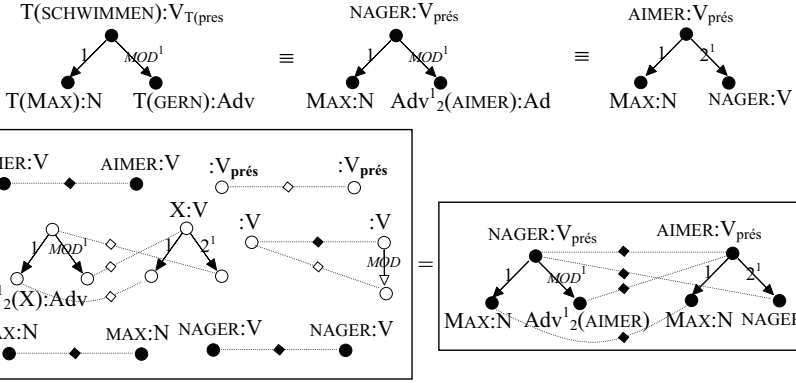


Fig. 4. La traduction (4)

La Fig. 4 montre le fonctionnement complet du système sur la traduction (3) : la première étape utilise uniquement le lexique bilingue sans modifier la structure ; la deuxième étape (détaillée dans la partie basse de la figure) effectue la restructuration. On notera qu'en plus de la règle de restructuration associée à Adv^1_2 et des règles de copie une règle permet de créer un lien entre deux verbes si l'un des deux est associé à un modifieur de l'autre. Cette règle est essentielle pour permettre la copie du temps présent de NAGER sur AIMER. Une telle règle revient à autoriser un changement de tête pendant la traduction : le sens lexicalisé comme tête de la phrase n'est plus le même[9].

Comme on le voit PUG permet de modéliser très simplement la combinaison des règles pour une simple fusion des objets noir et blanc (des nœuds, comme des liens de correspondance) et offre ainsi un moyen élégant d'écrire des règles de transfert.

4 Restructurations multiples

Une des propriétés les plus intéressantes du formalisme est qu'il permet de prendre en compte des restructurations multiples simultanées. Nous allons montre ça sur un exemple monolingue (ce qui ne change rien à la question) :

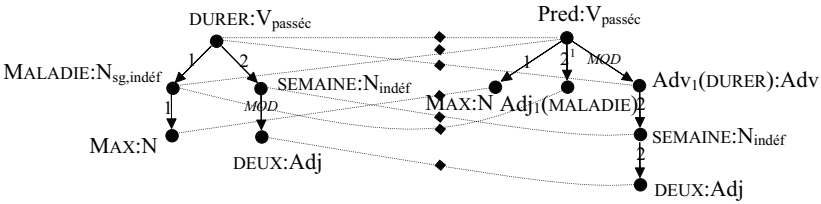
(5) **a.** *Max a été malade pendant 2 semaines*.
b. *La maladie de Max a duré deux semaines*.

[9] Il faudrait indiquer que ce lien est de nature « secondaire », bloquant ainsi le transfert d'actant entre les deux têtes. Le formalisme GUP permet aisément ceci en typant différemment les deux liens, à la différence des grammaires synchrones (Nesson *et al.* 2006) où les liens sont encodés par une coïndexation.

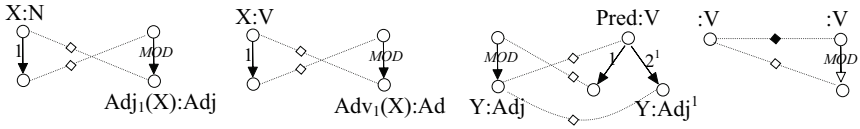
Pour réaliser, ce paraphrasage deux règles lexicales sont nécessaires :

MALADE = Adj_1(MALADIE)
PENDANT = Adv_1(DURER)

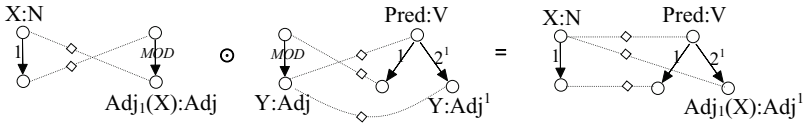
La fonction Adj_1 indique que MALADE est un Adj qui a le même sens que MALADIE et dont le gouverneur est le 1er actant de MALADIE. Idem pour Adv_1 (le 2ème actant de DURER devient le 2ème actant de PENDANT, lequel est une préposition, c'est-à-dire un averbe « transitif »). Par ailleurs, nous introduisons la fonction lexicale Pred, qui associe à chaque Adj un verbe support, lequel est la copule pour tous les Adj en français : Pred = ÊTRE. La paraphrase est réalisée par l'équivalence suivante :



Notons enfin que les règles de transfert doivent s'appliquer simultanément : il n'est pas possible de changer DURER en PENDANT puis MALADIE en MALADE, car en même temps que DURER devient PENDANT, son premier actant qui est un nom (MALADIE) doit devenir un verbe (en fait il devient un adjectif, MALADE, translaté en verbe par Pred). Pour ce transfert, quatre règles de restructuration vont devoir être agir conjointement. Les trois premières sont les règles structurelles associées aux fonctions lexicales Adj_1, Adv_1 et Pred. Les règles pour Adj_1 et Adv_1 disent que ce sont des modifieurs dont le gouverneur correspond au 1er actant du mot-clé ; la règle pour Pred associe à un Adj une tournure verbale dont le 1er actant est le gouverneur de l'Adj. La dernière est la règle de changement de tête vu précédemment[10].



Les règles associées à Adj_1 et Pred ne vont pas s'appliquer telle quelle. Elles doivent d'abord être combinées entre elle pour donner une nouvelle règle permettant le passage direct du N MALADIE à la structure Pred+Adj ÊTRE MALADE.



Les règles structurelles et les fonctions lexicales forment ainsi un système algébrique avec une opération de composition, que nous notons ⊙ (Kahane & Polguère 2001). La composition

[10] En changeant de tête syntaxique, on modifie fortement la structure informationnelle de la phrase, la partition sujet-verbe correspondant souvent en français à la partition thème-rhème. Une autre raison de la différence sensible de sens entre (5)a et b est le fait que le remplacement d'un adjectif par un nom entraîne l'introduction d'un nombre et d'une détermination sur ce nom et donc de deux nouvelles unités significatives que sont le singulier et le défini.

des règles structurelles est assurée par le « recollement » de la partie gauche de la 1ère règle sur la 2ème, comme on le voit sur la figure précédente.

5 Verbes de mouvement

Les verbes de mouvement représentent un problème classique de traduction :

(6) **a.** *Zoe swam accross the river.*
b. *Zoé a traversé la rivière à la nage.*

(7) **a.** *Zoe drove to the hospital.*
b. *Zoé est allé à l'hôpital en voiture.*

(8) **a.** *Max crawled out of the den.*
b. *Max est sorti de la tanière en rampant.*

Comme on le sait, pour reprendre la terminologie de l'article fondateur de Talmy 1976, l'anglais est *satellite-frame*, tandis que le français est *verb-frame*, c'est-à-dire que l'anglais indique la trajectoire du mouvement par un satellite (et la « manière » par le verbe), tandis que le français l'indique par le verbe principal (et la « manière » par un circonstanciel). Ceci se traduit lexicalement par le fait qu'un verbe comme SWIM ou CRAWL, à la différence de ses équivalents français, NAGER et RAMPER, possède un 2ème actant qui est un complément locatif. En conséquence, lorsque ce 2ème actant est instancié, le verbe anglais ne peut généralement pas être traduit par un verbe français et une restructuration est nécessaire pour attribuer ce 2ème actant à un autre verbe. C'est ce que fait la règle de la Fig. 5, qui remplace la préposition locative Y par la tournure verbale GO Y en même temps qu'elle remplace le verbe de mouvement X par un Adv^1_+ (le fait que X soit un verbe de mouvement assure que l'introduction de GO n'ajoute pas de sens). Un Adv^1_+ peut être exprimé par le gérondif (Adv^1_+(RAMPER) = *en rampant*). Le + indique que la relation *MOD* est additionnelle : il n'y a pas de relation prédicative entre un Adv_+ et son gouverneur ; les deux forment plutôt une coprédication. Certains verbes de mouvement possèdent un Adv^1_+ spécifique : Adv^1_+(NAGER) = À LA NAGE , Adv^1_+(CONDUIRE) = EN VOITURE.

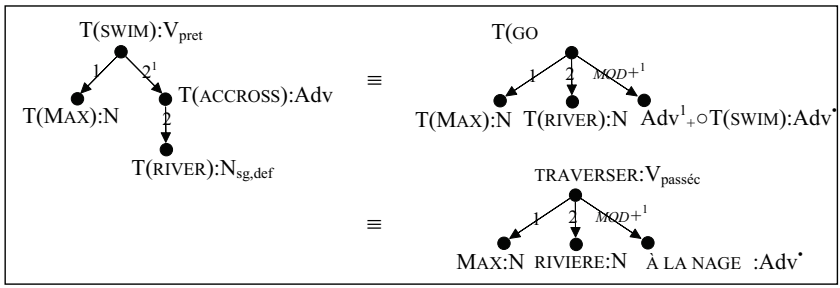


Fig. 5. La traduction (6) et la règle de restructuration correspondante.

Dorr 1994 argumente qu'u et propose une solution utilisant une représentation plus profonde supposée être une interlingua. Nasr *et al.* 1997 propose une solution avec transfert basée décomposition sous-jacente des UL

et mettant en jeu un échange des traits sémantiques correspondant à la trajectoire et la manière du mouvement. Notre solution est plus simple et notre lexique bilingue contient simplement des entrées telles que (GO OUT:V, SORTIR:V), (GO ACROSS:V, TRAVERSER:V), (GO TO:V, ALLER (à/jusqu'à):V) et la règle de restructuration de la Fig. 5 permet à ces règles de s'appliquer dans tous les cas.

6 Conclusion

L'architecture que nous avons présentée (et qui, rappelons-le, a été proposée dès les années 60) est particulièrement économique : la plus grande partie des règles utilisées appartiennent aux modèles monolingues (en particulier la description des collocations), les restructurations sont assurées par des règles universelles et le lexique bilingue est ainsi ramené à une taille minimale. Chaque règle de restructuration est ancrée par une fonction lexicale syntaxique qui contrôle quand la règle peut être utilisée, c'est-à-dire quand la règle de restructuration peut se combiner avec une correspondance entre deux UL modélisée par cette fonction lexicale dans les lexiques monolingues ou bilingue. Le formalisme des grammaires de correspondance polarisées permet une écriture élégante des règles et un contrôle facile de ce que manipule exactement chaque règle. L'implémentation du modèle doit faire l'objet d'études supplémentaires.

Notons qu'il est possible d'effectuer le transfert directement au niveau SyntS (voir par ex. Schubert 1987). Mais dans ce cas, on devra se faire correspondre des portions des SSyntS, portions qui correspondent en fait à des US avec leur régime. Un tel mécanisme (équivalent aux grammaires synchrones à la Ding & Palmer 2005 ou Nesson *et al.* 2006) revient donc à intégrer dans le système de transfert une description de la SyntS des US et rend le développement du lexique bilingue extrêmement plus coûteux (d'où le recours à une extraction automatique)[11].

Remerciements

Je remercie Laurence Danlos, Igor Mel'čuk et Alain Polguère et les trois relecteurs pour leurs commentaires sur la version préliminaire de cet article, ainsi que Leo Wanner et Igor Mel'čuk à nouveau pour les discussions acharnées à propos de leur travail sur le même sujet.

Références

ABEILLÉ A., SCHABES Y., JOSHI A. (1990) Using Lexicalized TAGs fo Machine Translation, *Proceedings of COLING,* Helsinki, Finland, 1-6.

APRESIAN J., BOGUSLAVSKY I., IOMDIN L., LAZURSKY A., SANNIKOV V., SIZOV V. & TSINMAN L. (2003) ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT, *Proceeding of MTT,* Paris, 279-288.

ARNOLD D., DES TOMBES L. (1987) Basic theory and methodology in Eurotra. In S. Nirenburg (ed.), *Machine Translation: Theoretical and Methodological Issue*, Cambridge University Press, 114-135.

DING Y., PALMER M. (2005) Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars, Actes de *ACL*, Ann Arbor.

[11] La question du recours à un modèle statistique, elle, n'est pas remise en cause. Des informations sur la fréquence (et les conditions d'utilisation) des US sont nécessaires pour le meilleur ajustement des traductions.

DANLOS D., SAMVELIAN P. (1992) Translation of the predicative elements of a sentence: category switching and aspect, in *Proceedings of Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, Canada.

DORR B. J. (1994) Machine Translation Divergences: A Formal Description and Proposed Solution, *Computational Linguistics*, 20(4):597-633.

DUCROT O. (1995) Les unités significatives. In Ducrot O. & J.-M. Schaeffer, *Nouveau dictionnaire encyclopédique des sciences du langage*, Paris : Seuil, 358-365.

GOLDBERG A. (1995) *Constructions. A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

KAHANE S. (2004) Grammaires d'Unification Polarisées, Actes de *TALN*, Fès, 10 p.

KAHANE S., LAREAU F. (2005), Grammaire d'Unification Sens-Texte : modularité et polarisation, Actes de *TALN*, Dourdan, 23-32.

KAHANE S., POLGUÈRE A. (2001) Formal foundations of lexical functions, Actes du *Workshop on Collocation, ACL,* Toulouse, 8 p.

KULAGINA O.S., MEL'CUK I.A. (1997) Automatic translation: some theoretical aspects and the design of a translation system, in A.D Booth (ed.) *Machine translation*, Amsterdam, 131-171.

LAVOIE B., KITTREDGE R. KORELSKY T., RAMBOW O. (2000) A Framework for MT and Multilingual NLG Systems Based on Uniform Lexico-Structural Processing, *Proc. of ANLP/NAACL 2000*, Seattle.

LINDOP J., TSUJII. J. (1993) Complex Transfer in MT: A Survey of Examples. Technical report, num 91, 5, Center for Computational Linguistics, Manchester, UMIST.

MARTINET A. (1960) *Éléments de linguistique générale*, Paris.

MEL'CUK I. (1988a) *Dependency Syntax: Theory and Practice*, SUNY.

MEL'ČUK I. (1988b) Paraphrase et lexique dans la théorie linguistique Sens-Texte — Vingt ans après, *Revue internationale de lexicologie et lexicographie*, Vol. 52/53, pp. 5-50/5-53.

MEL'CUK I, WANNER L. (2006) Syntactic mismatches in machine translation, *Machine Translation*, 20:2, 81-138.

NASR A., RAMBOW O., PALMER M., ROSENZWEIG J. (1997) Enriching Lexical Transfer With Cross-Linguistic Semantic Features. In *Proceedings of the AMTA/SIG-IL 1st Workshop on Interlingua*, San Diego, CA. Published as New Mexico State University Computing Research Laboratory technical report MCCS-97-314.

NESSON R., SHIEBER S., RUSH A. (2006) Induction of probabilistic synchronous tree-insertion grammars for machine translation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts.

SCHUBERT K. (1987) *Metataxis: Contrastive Dependency Syntax for Machine Translation*. Dordrecht/Providence: Foris

TALMY L. (1976) Semantic Causative Types. In M. Shibitani (ed.) *Syntax and Semantics 6: The Grammar of Causative Constructions,* Academic Press : N.Y, 43-116.

TESNIÈRE L. (1959) *Éléments de syntaxe structurale*, Paris: Klincksieck.

VIEGAS E. (1997) Mismatches and divergences: the continuum perspective, *Proc. of TMI*, Santa Fe.

ŽOLKOVSKIJ A., MEL'ČUK I. (1967) O semantičeskom sinteze, *Problemy Kybernetiki, 19,* 177-238. [trad. franç. : 1970, *T.A. Information, 2*, 1-85.]

Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF - ISO 24613

Aïda KHEMAKHEM[1], Bilel GARGOURI[1],
Abdelhamid ABDELWAHED[2], Gil FRANCOPOULO[3]

[1] Laboratoire MIRACL, FSEG-SFAX B.P. 1088, 3018 SFAX – TUNISIE
[2] Unité de recherche LSCA, FLSH-SFAX B.P. 553, 3018 SFAX – TUNISIE,
[3] INRIA-Loria
khemakhem.aida@gnet.tn, Bilel.Gargouri@fsegs.rnu.tn
abdelhamid.abdelwahed@yahoo.fr, gil.francopoulo@wanadoo.fr

Résumé. Dans cet article, nous spécifions les paradigmes de flexion des verbes arabes en respectant la version 9 de LMF (Lexical Markup Framework), future norme ISO 24613 qui traite de la standardisation des bases lexicales. La spécification de ces paradigmes se fonde sur une combinaison des racines et des schèmes. En particulier, nous mettons en relief les terminaisons de racines sensibles aux ajouts de suffixes et ce, afin de couvrir les situations non considérées dans les travaux existants. L'élaboration des paradigmes de flexion verbale que nous proposons est une description en intension d'*ArabicLDB* (Arabic Lexical DataBase) qui est une base lexicale normalisée pour la langue arabe. Nos travaux sont illustrés par la réalisation d'un conjugueur des verbes arabes à partir d'*ArabicLDB*.

Abstract. In this paper, we specify the inflected paradigms of Arabic verbs with respect to the version 9 of LMF (Lexical Markup Framework) which is the expected ISO 24613 standard dealing with the standardisation of lexical databases. The specification of these paradigms is based on a combination of schemes and roots. In particular, we highlight the role of root endings that is not considered in other researches and that may generate erroneous forms while concatenating suffixes. The development of verbal inflected paradigms that we propose is an intentional component of *ArabicLDB* (Arabic Lexical DataBase) which is a normalized Arabic lexical database that we developed according to LMF. Our works are illustrated by the realization of a conjugation tool for Arabic verbs using *ArabicLDB*.

Mots-clés : langue arabe, paradigmes de flexion verbale, base lexicale, norme ISO 24613, LMF, lexical markup framework, conjugueur des verbes arabes.

Keywords: Arabic, inflected paradigms of verb, lexical database, norm ISO 24613, LMF, lexical markup framework, conjugation of arabic verbs.

1 Introduction

L'arabe est une langue à la fois dérivationnelle et flexionnelle (Blachère et al, 1975). Elle se caractérise par une morphologie assez complexe à la manière de la conjugaison de ses verbes

qui génère plusieurs formes fléchies variant d'un verbe à un autre. Ainsi, s'accroît le besoin de modéliser les paradigmes de flexion de l'arabe afin de limiter l'espace de représentation des formes fléchies lors de la construction d'une base lexicale.

Dans le présent papier, nous nous intéressons à la modélisation des paradigmes de flexion des verbes arabes en vue de construire une description en intension des verbes dans une base lexicale. Nos investigations sont fondées sur les principaux travaux existants dans ce domaine (El-Dahdah, 1999), (Ammar et al, 1999), (Abdelwaheh, 1996) et (Ibn elqataa, 2003).

Ces travaux classent les verbes selon des critères qui se fondent sur les consonnes sensibles à la modification des voyelles voisines. Néanmoins, nous avons pris en considération certains cas liés à l'ajout d'un suffixe qui peuvent engendrer des formes erronées. Partant de cette constatation, nous proposons d'adapter la classification des racines selon les besoins du traitement automatique en vue de couvrir des cas omis dans la quasi-totalité des travaux sur le sujet.

En ce qui concerne la modélisation et la réalisation, nous profitons du travail en cours au sein de l'ISO de la spécification LMF (Lexical Markup Framework) (Francopoulo, 2003) qui propose une représentation lexicale standard pour les langues les plus utilisées tout en couvrant, entre autres, la représentation des paradigmes de flexion. Ce projet est en cours de validation par le comité TC37/SC4 de l'ISO sous la référence ISO-24613 (Francopoulo, 2006). Comme illustration de l'application de LMF sur le cas de la langue arabe, nous présentons la base *ArabicLDB* qui se limite dans sa version actuelle à la représentation du niveau morphologique (Khemakhem et al, 2006), (Khemakhem, 2006).

Le présent travail est réalisé selon les recommandations de la version 9 de LMF appliquées à la base *ArabicLDB* représentant une description des paradigmes de flexion verbale.

Dans la première partie, nous présentons les travaux actuels de modélisation des paradigmes de flexion des verbes arabes. Ensuite, nous donnerons une idée sur le projet de normalisation LMF, en s'intéressant en particulier à l'extension des paradigmes de flexion. Dans la section suivante, nous présenterons la modélisation des paradigmes verbaux de la langue arabe selon LMF. Puis nous spécifierons les verbes types de la langue arabe. Enfin, nous terminerons par l'implémentation des paradigmes de la base ArabicLDB avec un descriptif du conjugueur réalisé.

2 Aperçu sur les travaux de modélisation des paradigmes des verbes arabes

2.1 Caractéristiques des verbes arabes

Dans la langue arabe, il y a plus de 16 000 verbes. Un verbe peut avoir 109 formes fléchies quand il admet à la fois la voix active et passive et 57 formes fléchies (FF) lorsqu'il n'admet que la voix active (Ammar et al, 1999). En général, les verbes sont des formes dérivées à partir d'une racine et d'un schème. La racine est purement consonantique qui peut être formée soit par trois consonnes formant des verbes trilitères, soit par quatre consonnes formant des verbes quadrilitères. Le schème peut être considéré comme une représentation formelle constituée par 3 ou 4 consonnes (ف, ع, ل) qui sont totalement vocalisées, ou comme un moule sur laquelle coule la racine (Baloul, 2003). En totalité, il y a 19 schèmes verbaux qui peuvent être soit nus, soit augmentés dérivant de trois consonnes de la racine par modification des voyelles, par redoublement de la deuxième lettre de la racine, par adjonction et même par

intercalation d'affixes (préfixe, infixe, suffixe). Les verbes augmentés se conjuguent avec les mêmes préfixes et suffixes que le verbe sans augment. De ce fait, une racine peut générer au maximum 19 verbes et les schèmes correspondants peuvent donner 22 modèles de conjugaison différents. En effet, il y a le schème فعَلَ [fa'ala] qui peut avoir trois alternances différentes de conjugaison selon la nature de la voyelle de la 2e consonne de la racine de l'inaccompli يَفْعُلُ [yaf'ulu], يَفْعِلُ [yaf'ilu], et لَيَفْعَ [yaf'alu]. En plus, le schème فعِلَ [fa'ila] peut donner deux alternances différentes de conjugaison pour la même raison (El-Dahdah, 1999).

La conjugaison d'un verbe arabe consiste à engendrer l'ensemble des formes que peut prendre ce verbe à la voix active ou passive. Ces formes varient selon le mode et l'aspect (accompli, inaccompli indicatif, inaccompli subjonctif, inaccompli apocopé et l'impératif), comme elles varies selon le nombre (singulier, duel, pluriel) et le genre (masculin, féminin). Elles varient aussi selon la personne ou les personnes représentées par le sujet : تُكَتَبْ[katab**tu**], تُمَاكَتَبْ [katb**tumâ**]…

Généralement, les règles de conjugaison d'un verbe arabe peuvent engendrer les modifications suivantes : le changement des voyelles, l'ajout d'un préfixe ou d'un suffixe, et la suppression d'un *hamza wasliya* au début ou d'une voyelle à la fin. De plus, ces modifications peuvent avoir recours à des règles phonologiques qui opèrent sur certaines formes verbales particulières.

2.2 Les principales propositions du domaine

Les travaux sur la langues arabe sont très nombreux mais très dispersés comparés à ceux d'autres langues comme l'anglais ou le français. Parmi les principaux travaux sur les verbes arabes, nous citons « **معجم تصريف الأفعال العربيّة** » (El-Dahdah, 1999)**,** « **Les verbes arabes** » (Ammar et al, 1999), « **قِرَاءة في التصريف العربي .بنية الفِعل** » (Abdelwaheb, 1996) et «**باتك لاعفألا** » (Ibn elqataa, 2003) qui traitent un nombre important de verbes, mais qui sont destinés à une utilisation manuelle.

Signalons que les travaux sur les verbes possèdent quelques points communs, comme la distribution des verbes et la détermination de leur nature, ainsi que l'idée principale qui consiste à traiter les règles phonologiques comme des règles de conjugaison ou morphologique. Selon ces auteurs, les causes de déclenchement des règles phonologiques sont limitées par la nature des consonnes de la racine. Ce qui a pour effet de déterminer un classement des paradigmes de flexion et des verbes. Selon cette optique, deux principales classes de verbe sont distinguées: les verbes sains (حيحص) qui ne comportent pas de lettres défectueuses, et les verbes défectueux (معتلّ) qui contiennent une ou deux lettres défectueuses causant des altérations importantes au cours de la conjugaison. Un verbe sain peut contenir la lettre *hamza* ou *šadda* qui peut engendrer des conjugaisons irrégulières.

Bien qu'il y ait des points communs, les différences impliquent des conséquences importantes pour la représentation et la couverture : Ammar traite seulement 10 000 verbes, alors que El-Dahdah traite plus de 16 000 des verbes arabes. Notons que ce dernier auteur couvre la plupart des verbes quadrilitères ainsi que les verbes trilitères rarement utilisés.

La complexité morphologique de la langue arabe est constatée entre autres dans l'utilisation des règles phonologiques au cours du calcul d'une forme fléchie. Par conséquent, les linguistes ont proposé la classification des verbes selon la racine et ils ont valorisé les consonnes (la lettre *hamza,* les lettres défectueuses ou le signe *šadda*) sensibles aux changements de leurs voyelles (précédente et suivante). Cependant, l'automatisation directe

de cette classification, peut nous donner des formes erronées. Pour éviter l'apparition de ces formes, il faut valoriser les consonnes sensibles à l'ajout d'un suffixe (les terminaisons qui peuvent être fusionnées avec la première consonne du suffixe) (Khemakhem, 2006).

Lemme	حَزَنَ	[ḥazana]
Forme Fléchie intermédiaire avec هُنَّ	حَزَنْنَ	
Forme Fléchie finale avec هُنَّ	حَزَنَّ	[hazanna]

Figure 1 : Exemple de fusion d'une terminaison et suffixe utilisant une règle phonologique

Dans cet exemple, si ce verbe est traité comme les autres verbes sains, la forme intermédiaire « حَزَنْنَ » [hazanna] sera une forme erronée qui nécessitera l'intervention d'une règle phonologique pour donner la forme finale « حَزَنَّ » [hazanna].

3 LMF : future norme ISO 24613

3.1 Présentation générale

LMF (Lexical Markup Framework) est un projet en cours de validation par le sous comité TC-37/SC-4 de l'ISO sous la référence ISO-24613. Le projet traite de la normalisation des bases lexicales à large couverture (Francopoulo et al, 2006). Ce projet est né du besoin exprimé par les différentes délégations nationales de pouvoir représenter, échanger et fusionner les bases lexicales actuelles et futures. Notons que le besoin de fusion est celui qui est le plus souvent exprimé. La totalité de LMF est spécifiée à l'aide d'UML (Unified Modeling Language) qui s'impose maintenant comme le standard de modélisation des structures. Le format d'échange est défini en XML.

Un lexique LMF doit comporter une partie noyau et un ou plusieurs modules d'extension qui sont spécifiés par des "packages" UML. Le noyau traite des notions de lexique (en tant que conteneur), de mot, de forme et de sens. Le noyau est à considérer à la fois comme le squelette du lexique et comme le dénominateur commun à plusieurs lexiques, ce qui est essentiel pour les opérations de fusion. Le noyau est obligatoire. Autour de ce dernier, le créateur du lexique peut définir des modules d'extension pour la morphologie, les paradigmes de flexion, la syntaxe, la sémantique et/ou les notations multilingues.

LMF définit la structure du lexique mais n'en précise pas les détails. La structure est décorée par des couples attribut-valeur à prendre dans un registre de catégories de données (RCD) qui est géré par l'ISO dans le cadre de l'ISO-12620 (http://syntax.inist.fr/) (Romary et al, 2003).

Ainsi, LMF permet deux axes de variation : d'une part le choix des modules d'extension et d'autre part le choix des catégories de données. Par exemple, un gestionnaire de lexique de morphologie n'aura pas à s'embarrasser des mécanismes descriptifs de la syntaxe, s'il n'en éprouve pas le besoin. De même, ce même gestionnaire sélectionnera les valeurs /masculine/ et /feminine/ pour le français alors qu'il choisira /masculine/, /feminine/ et /neuter/ pour l'allemand.

3.2 Le module d'extension des paradigmes de flexion

Le module des paradigmes proposé dans LMF est une description simple des modifications qui sont apportées au lemme pour calculer toutes les formes fléchies d'une entrée lexicale.

Cette description utilise huit opérations (/add/, /addBefore/, /addAfter/, /remove/, /removeAfter/, /removeBefore/, /substitute/, /copy/) qui sont liées à un calculateur de forme fléchie spécifique à une combinaison de traits morphologiques (Francopoulo et al, 2006).

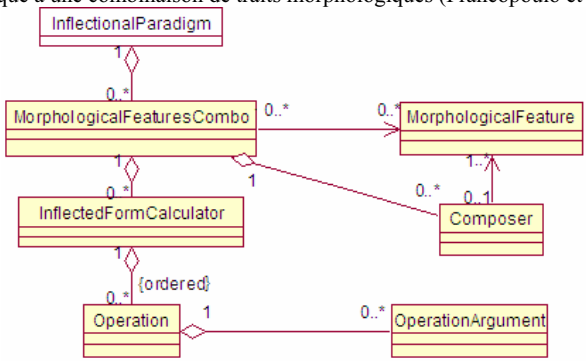


Figure 2 : L'extension des paradigmes de flexion selon LMF

Les paradigmes de flexion comportent plusieurs instances de la classe *Morphological FeaturesCombinator*. Ces dernières sont relatives à une classe abstraite qui fait le lien entre des traits morphologiques et un ou plusieurs *InflectedFormCalculators* qui regroupent des *Operations* et des *OperationArguments* permettant de calculer les FF correspondantes.

4 La base lexicale ArabicLDB

ArabicLDB (Khemakhem, 2006) est une nouvelle base lexicale arabe, qui est conforme à la révision 9 de la future norme LMF. Elle est implémentée en XML en respectant la DTD proposée dans cette révision (Francopoulo et al, 2006). Le choix a été pris de décrire explicitement toutes les FF des noms et de décrire les verbes en intension à l'aide des paradigmes de flexion.

Les entrées lexicales d'*ArabicLDB* peuvent être des formes dérivées (nominales ou verbales), des noms non dérivés et des particules, qui sont totalement vocalisées. Pour chaque partie du discours, un ensemble de catégories de données est sélectionné à partir du RCD quand c'était possible. Certaines valeurs sont absentes du registre actuel. Nous les avons utilisées et testés, puis, nous avons demandé leur ajout dans le registre de l'ISO.

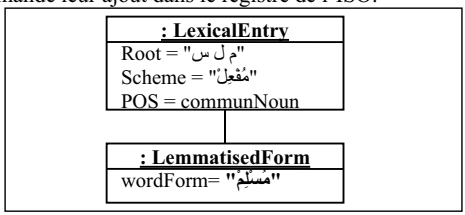


Figure 3 : Exemple d'une entrée lexicale

Les entrées lexicales portent un identifiant et l'information de partie du discours. Ce choix implique la création de deux entrées pour des formes identiques à parties de discours distinctes (كَمْ : pronom interrogatif vs. كَمْ : pronom allusif (كِنَايَة)). En contrepartie, des

homonymes à partie de discours identique (سَاعَة : une heure vs. سَاعَة : une montre) ne sont pas traités dans deux entrées différentes.

Pour la langue arabe, le verbe et la plupart des noms sont identifiés par une racine et un schème. En effet, nous utilisons /root/ et /scheme/ comme des couples attribut-valeur de la classe *LexicalEntry*, qui est la classe permettant de représenter la notion de mot.

Actuellement, nous avons un outil d'alimentation de cette base et d'autres outils pour l'exploitation. Nous avons des modules d'acquisition des paradigmes (256) et des entrées lexicales (plus de 16 000 verbe, 500 particules…). Pour l'exploitation, nous avons des modules de recherche et de conjugaison des verbes.

5 Les paradigmes de flexion verbale de l'arabe selon LMF

Les paradigmes de flexion sont des représentations communes à un grand nombre de mots dans une optique de description en intention. Ils factorisent la connaissance linguistique sous forme d'un prototype de conjugaison. Ils facilitent donc la maintenance et réduisent la taille de la base de données. Cependant, le chevauchement entre les différents niveaux linguistiques dans la langue arabe, notamment morphologique et phonologique, et l'absence de la phonologie dans la norme, montrent que ces paradigmes nécessitent une investigation très détaillée. En effet, nous suivons une démarche qui procède en trois étapes: la classification, l'identification des verbes types et l'application des opérations (selon LMF).

5.1 Classification des verbes

L'objectif de cette phase était l'étude des consonnes qui déterminent une règle phonologique. Après l'étude des classifications des racines déjà proposées dans la littérature, nous les avons précisé en valorisant les terminaisons (les lettres ت et ن) qui peuvent se fusionner avec la première consonne du suffixe. La Figure 4 donnée ci-après, illustre la considération des terminaisons lors de la classification des verbes. Dans cette figure, la sous classe des verbes sains (سَالِم) sera constituée de trois sous classes : la première (*) regroupe les verbes qui ne se terminent ni par la lettre ت ni par la lettre ن, la deuxième (#ن) regroupe les verbes ayant la terminaison ن et la troisième (#ت) regroupe les verbes ayant la terminaison ت.

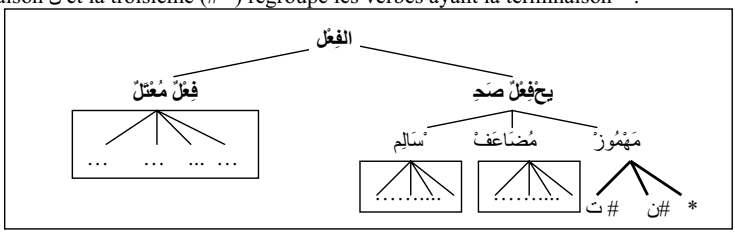


Figure 4 : Rôle des terminaisons dans la classification des verbes

En résultat, nous avons dégagé 42 classes de racines verbales en considérant les terminaisons qui peuvent se fusionner avec la première consonne du suffixe.

5.2 Identification des verbes types

Le modèle LMF n'a pas d'extension phonologique, ce qui nous amène à traiter les transformations phonologiques comme des flexions morphologiques. Et nous n'avons pas que 22 modèles de conjugaison puisqu'il faut considérer les différents types de racine possibles.

Par ailleurs, nous avons combiné les modèles de conjugaison (les 22 classes de schème) et les classes de racine. Ensuite, nous avons optimisé les résultats pour éviter la redondance des paradigmes et l'apparition de paradigmes superflus: il y a des schèmes qui revêtent un sens assez particulier et qui sont utilisés uniquement avec quelques classes de racine.

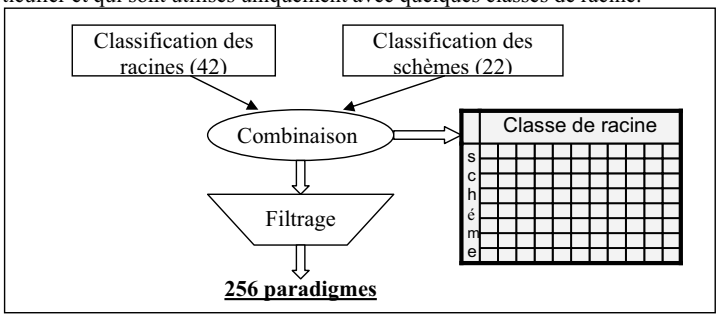


Figure 5 : Démarche de préparation des paradigmes verbaux

Après la combinaison et l'optimisation des combinaisons des classes de racine et des classes de schème, nous avons trouvé 256 paradigmes qui correspondrent aux verbes types (Khemakhem, 2006).

5.3 Application des opérations proposées par LMF pour le cas de l'arabe

Dans cette section, il s'agit d'étudier la couverture de toutes les transformations possibles au cours de la conjugaison des verbes par les opérations proposées par LMF.

Nous rappelons que les règles de conjugaison d'un verbe arabe peuvent engendrer les modifications suivantes : le changement des voyelles, l'ajout d'un préfixe ou d'un suffixe, et la suppression d'un hamza wasliya au début ou une voyelle à la fin. Nous avons utilisé les opérations proposées dans LMF pour appliquer les règles de conjugaison de l'arabe. Comme illustration, nous présentons dans la Figure 6, l'exemple de calcul de la forme fléchie « يَسْتَقْبِلُ » [yastaqbilu] à partir du lemme « اِسْتَقْبَلَ » [istaqbala].
Nous avons encore des cas particuliers qui nécessitent des modifications particulières liées à l'application des règles phonologiques. Ces modifications sont :

- la transformation de la graphie de *hamza*, ce qui nécessite parfois les opérations de remplacement **substitute** ou de suppression **remove**.
- la transformation des lettres défectueuses, ce qui nécessite parfois les opérations de remplacement **substitute** ou de suppression **remove**.
- la transformation de *šadda* qui peut être supprimé par **remove**, la consonne impliquée dupliquée par **copy.** De plus, des voyelles peuvent être ajoutées par **add**.
- la fusion du suffixe avec la dernière consonne de la racine ن ou ت. Elle est assurée par la suppression **remove** d'une consonne avec sa voyelle et l'ajout **add** de *šadda*.

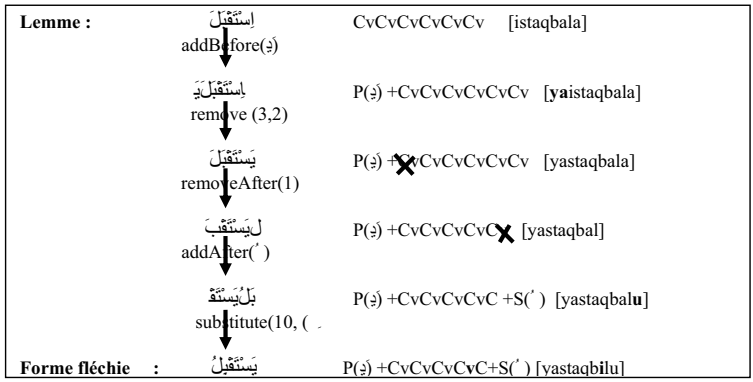


Figure 6 : Exemple d'un calcul d'une forme fléchie

D'après l'étude de tous les cas particuliers liés à la phonologie, nous pouvons conclure que les opérations proposées par la norme LMF couvrent toutes les flexions de la langue arabe.

6 Elaboration de la base des paradigmes

6.1 La base des paradigmes

Cette base comporte la description en XML des combinaisons des traits morphologiques pour chaque verbe type. Une interface appropriée a été mise en place pour faciliter la création d'un paradigme de flexion. L'utilisateur peut choisir une combinaison et spécifier les opérations en précisant les arguments associés. Comme illustration, nous présentons, dans la Figure 7, l'exemple de « يَسْتَقْبِلُ» [yastaqbilu] de la figure précédente.

```
<Lexicon> <InflectionalParadigm id="asIstaqbala">
<MorphologicalFeaturesCombo><MorphologicalFeature att="verbFormAspect" val="unaccomplished"/>
                            <MorphologicalFeature att="verbFormMood" val="indicative" />
                            <MorphologicalFeature att="voice" val="activeVoice" />
                            <MorphologicalFeature att="person" val="thirdPerson" />
                            <MorphologicalFeature att="grammaticalNumber" val="singular" />
                            <MorphologicalFeature att="grammaticalGender" val="masculine" />
   <InflectedFormCalculator>  <DC att="stem" val="0" />
     <Operation> <DC att="graphicalOperator" val="addBefore" />
         <OperationArgument> <DC att="chain" val="يَ" /> </OperationArgument>  </Operation>
      <Operation> <DC att="graphicalOperator" val="remove" />
         <OperationArgument> <DC att="pos" val="3" />
         <OperationArgument> <DC att="number" val="2" /> </OperationArgument>  </Operation>
      <Operation> <DC att="graphicalOperator" val="removeAfter" />
         <OperationArgument>  <DC att="number" val="1" /> </OperationArgument>  </Operation>
      <Operation> <DC att="graphicalOperator" val="addAfter" />
         <OperationArgument> <DC att="chain" val="ُ" /> </OperationArgument> </Operation>
      <Operation> <DC att="graphicalOperator" val="substitute" />
         <OperationArgument> <DC att="chain" val="ْ" /> </OperationArgument>
         <OperationArgument> <DC att="pos" val="10" /> </OperationArgument>  </Operation>
   </InflectedFormCalculator></MorphologicalFeaturesCombo>…
</InflectionalParadigm>…..</Lexicon>
```

Figure 7 : Combinaison de traits morphologiques du paradigme du verbe" اِسْتَقْبَلَ"

6.2 Le module d'extension morphologique de la base ArabicLDB

Après l'élaboration des paradigmes qui se caractérisent par un identifiant, nous pouvons spécifier pour chaque verbe son paradigme de flexion au niveau du lemme. En plus, nous ajoutons une catégorie de données /havePassive/ qui est spécifique pour les verbes. Elle prend **no** si le verbe admet seulement la voix active et **yes** si le verbe admet les deux voix.

```
<lexicalEntry>   <DC att="root" val="ب ت ك" />
                 <DC att="scheme" val="فعَلَ" />
                 <DC att="pos" val="verb" />
                 <lemmatisedForm paradigm="asKataba"> <DC att="wordForm" val="كَتَبَ" />
                                                      <DC att="havePassive" val="yes" />
                 </lemmatisedForm>
</lexicalEntry>
```

Figure 8 : Exemple d'une entrée lexicale verbale d'ArabicLDB

Dans cet exemple, nous avons spécifié le paradigme du verbe "كَتَبَ" [kataba] comme attribut de ce lemme. Cet attribut joue le rôle d'une référence pour trouver le paradigme convenable à un verbe dans le module des modes de flexion. Ce paradigme peut être appliqué totalement sur ce verbe puisque la catégorie de données /havePassive/ prend la valeur **yes**. Autrement dit, ce verbe admet les deux voix (active et passive).

7 Mise en œuvre d'un conjugueur

Le conjugueur de la base *ArabicLDB* utilise le module des paradigmes de flexions pour engendrer toutes les formes fléchies d'un verbe donné à partir du lemme. Dans un premier temps, le programme accède à l'extension morphologique pour chercher l'identifiant du paradigme du verbe. Dans un deuxième temps, il accède à l'extension des modes de flexion pour importer le paradigme en question. Nous rappelons que ce paradigme est composé de plusieurs combinaisons de traits morphologiques dont chacune d'elles possède des opérations et des arguments (position, chaîne, etc.) qui permettent le calcul des formes fléchies.

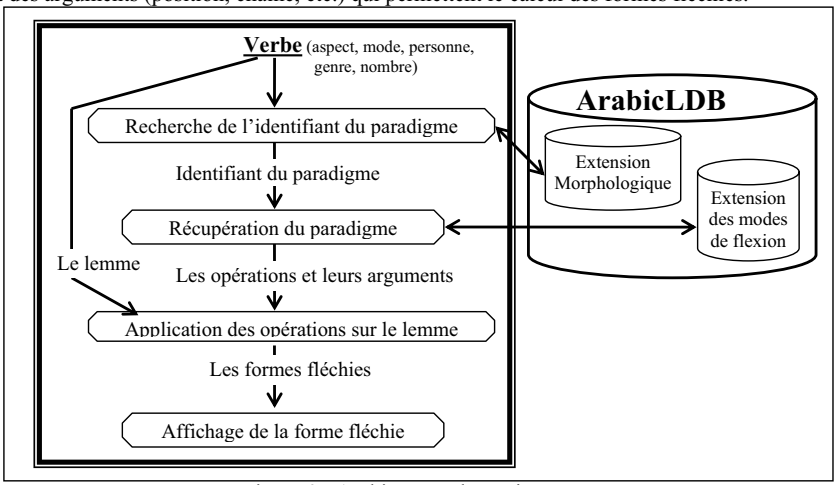


Figure 9 : Architecture du conjugueur

Notre conjugueur est composé de quatre modules :

- **Recherche de l'identifiant du paradigme** : c'est la première étape de ce conjugueur qui consiste à trouver l'identifiant du paradigme du verbe en question, en accédant à l'extension morphologique de la base *ArabicLDB*.
- **Récupération du paradigme** : cette étape utilise l'identifiant du paradigme pour accéder à l'extension des modes de flexion et importer les données de ce paradigme.
- **Application des opérations sur le lemme** : ce module permet la génération les formes fléchies en se basant sur les opérations et leurs argument.
- **Affichage de la forme fléchie** : nous affichons la forme fléchie générée avec ses traits morphologiques.

8 Conclusion

L'absence d'une représentation intensionnelle robuste pour les verbes arabes, nous a poussé à étudier ce domaine. En plus, l'apparition d'une norme ISO (LMF) nous a encouragé à élaborer les paradigmes de flexion des verbes arabes selon cette norme sachant que nous avons déjà utilisé LMF pour élaborer la base morphologique ArabicLDB.

En ce qui concerne l'identification des verbes types, nous avons adopté les classifications proposées dans les travaux existants afin de combiner les classes résultantes avec les classes de schème tout en mettant en relief les terminaisons des racines sensibles à l'ajout des suffixes. Ensuite, nous avons filtré ces combinaisons en utilisant des critères d'ordre sémantique pour aboutir à la spécification de 256 paradigmes verbaux. Ces paradigmes forment la représentation en intension des verbes dans la base *ArabicLDB* qui comporte 16 000 verbes.

Un conjugueur a été mis en place comme outil d'exploitation et de vérification en s'appuyant sur *ArabicLDB*. Ce conjugueur est utilisé dans un contexte d'enseignement de l'arabe.

Références

ABDELWAHED A. (1996).. كلية الآداب و العلوم الإنسانيّة بصفاقس، تونس.
AMMAR S., DICHY J. (1999). *Les verbes arabes* (Collection Bescherelle). Editions HATIER, Paris.
BALOUL S. (2003). Développement d'un système automatique de la parole à partir du texte arabe standard voyellé. Thèse de doctorat de l'université du MAINE, Le Mans, France.
BLACHERE R., GAUDEFROY-DEMOMBYNES M. (1975). *Grammaire de l'arabe classique.* Lieu : Edition Maisonneuvre-Larose, Paris.
EL-DAHDEH A. (1999). نانبل ،توريب ،نورشان نانبل قبتكم.
FRANCOPOULO G. (2003). Proposition de normalisation de norme des lexiques pour le traitement automatique du langage. INRIA/LORIA-ACTION SYNTAXE, Version-1.3.
FRANCOPOULO G., GEORGE M. (2006). ISO/TC 37/SC4 N130 Rev.9. Language resource management − Lexical markup framework (LMF).
IBN ELQATAA (2003). تـوريب . ةيملعلا بتكلا راد
KHEMAKHEM A., GARGOURI B., ABDELWAHED A. (2006). LMF est-il convenable pour la langue arabe ? Actes de *Journées sur le Traitement Automatique de la Langue Arabe JTALA'06*, Rabat, Maroc.
KHEMAKHEM A. (2006). "ArabicLDB : une base lexicale normalisée pour la langue arabe". Mémoire de master en Systèmes d'Information et Nouvelles Technologies. Lieu : FSEG, Sfax, Tunis. (http://www.tagmatica.fr/doc/MemoireAida.pdf).
ROMARY L., WRIGHT S., FARRA S., GILLAM L. (2003). ISO TC 37/SC4 N055, Language resource management - Implementing a data category registry within ISO TC37.

Du bruit, du silence et des ambiguïtés : que faire du TAL pour l'apprentissage des langues ?

Olivier KRAIF, Claude PONTON
LIDILEM - Laboratoire de linguistique et didactique des langues étrangères et maternelles (http://www.u-grenoble3.fr/lidilem/labo)
{Olivier.Kraif,Claude.Ponton}@u-grenoble3.fr

Résumé. Nous proposons une nouvelle approche pour l'intégration du TAL dans les systèmes d'apprentissage des langues assisté par ordinateur (ALAO), la stratégie « moins-disante ». Cette approche tire profit des technologies élémentaires mais fiables du TAL et insiste sur la nécessité de traitements modulaires et déclaratifs afin de faciliter la portabilité et la prise en main didactique des systèmes. Basé sur cette approche, ExoGen est un premier prototype pour la génération automatique d'activités lacunaires ou de lecture d'exemples. Il intègre un module de repérage et de description des réponses des apprenants fondé sur la comparaison entre réponse attendue et réponse donnée. L'analyse des différences graphiques, orthographiques et morphosyntaxiques permet un diagnostic des erreurs de type fautes d'orthographe, confusions, problèmes d'accord, de conjugaison, etc. La première évaluation d'ExoGen sur un extrait du corpus d'apprenants FRIDA produit des résultats prometteurs pour le développement de cette approche « moins-disante », et permet d'envisager un modèle d'analyse performant et généralisable à une grande variété d'activités.

Abstract. This paper presents the so-called "moins-disante" strategy, a new approach for NLP integrating in Computer Assisted Language Learning (CALL) systems. It is based on the implementation of basic but reliable NLP techniques, and put emphasis on declarative and modular processing, for the sake of portability and didactic implementation. Based on this approach, ExoGen is a prototype for generating activities such as gap filling exercises. It integrates a module for error detection and description, which checks learners' answers against expected ones. Through the analysis of graphic, orthographic and morphosyntactic differences, it is able to diagnose problems like spelling errors, lexical mix-up, error prone agreement, bad conjugations, etc. The first evaluation of ExoGen outputs, based on the FRIDA learner corpus, has yielded very promising results, paving the way for the development of an efficient and general model tailored to a wide variety of activities.

Mots-clés : ALAO, apprentissage des langues, diagnostic d'erreur, feed-back d'erreur.
Keywords: CALL, language learning, error diagnosis, error feedback.

1 Introduction

Dans le cadre de l'ALAO (Apprentissage des Langues Assisté par Ordinateur), et plus particulièrement pour les systèmes dits « de structure » (systèmes de répétitions, d'entraînement, tutoriels, etc. par opposition aux systèmes de référence et aux systèmes

d'exploration, cf. Wyatt, 1987; Meunier, 2000), la détection et l'analyse d'erreurs constituent un élément central pour un diagnostic et une production de rétroactions adaptées permettant ainsi un apprentissage interactif, autonome et personnalisé. Or, la plupart de ces systèmes actuels ne disposent que de techniques rudimentaires de tests d'identité de chaînes entre « réponse donnée » et « réponse attendue ». Ces approches ne peuvent conduire qu'à de simples rétroactions du type vrai/faux n'offrant pas ainsi à l'apprenant une possibilité de réflexion sur les stratégies qu'il a mises en œuvre et lui permettre ainsi de modifier sa production langagière. En traitant les différents niveaux de la langue, le TAL semble pouvoir offrir de meilleures perspectives à cette problématique. Toutefois, comme le fait remarquer J. Rézeau (2001), «(…) *on constate que la quasi-totalité des didacticiels de langue sur le marché à la fin des années 1990 proposent des exercices du premier type* (i.e. qui attendent une seule réponse)*, et donc une analyse de réponse que nous qualifierons de "minimale"* (i.e. vrai/faux)». Outre le manque de communication entre les chercheurs et praticiens des différents domaines, l'explication de cette désaffection du TAL en ALAO découle de diverses raisons liées aux contextes didactiques des systèmes et donc de l'analyse d'erreurs.

◆ **Analyse de textes libres** : Des systèmes comme FreeText (Granger et al., 2001) ou « le correcteur 101 didactique » de la société Machina Sapiens visent l'analyse d'erreurs sur des textes libres d'apprenants. Or, comme le reconnaît lui-même S. L'Haire, l'un des participants au projet FreeText, le système pèche par « une trop grande surdétection d'erreurs » (L'Haire, 2004). La même critique s'applique au correcteur 101 Didactique. Par exemple, l'analyse par ce logiciel d'un texte d'apprenant de niveau intermédiaire issu du corpus FRIDA (Granger et al., 2001) produit :

« Dans tout le monde, il y a plus que plusieurs langues étrangères [étranger]. Ce pour sa [ça], qui parler [parle] ou connaître plusieurs de langues est [sont] nécessaire. Mais [mai ?], je crois qui parler trois ou quattre langues sont sufficents, parce qui n'est pas possible étudier [étudie?] tout le [toutes les] langues. » (détection en souligné, correction proposée entre crochets)

Si la détection des mots inconnus (*quattre*, *sufficents*…) est relativement bien traitée, le traitement de l'homophonie indique fréquemment des erreurs inexistantes. Seuls les accords de courtes portées sont bien analysés et des erreurs grossières d'analyse syntaxique (* *plusieurs langues étranger*) compromettent une utilisation en autonomie complète.

◆ **Analyse de textes sous contrôle** : Pour éviter les travers d'une analyse de textes libres, certaines approches ont tenté de contrôler divers paramètres de la production. Par exemple, le système ALEXIA (Selva, Chanier, 2000) travaille sur un domaine ciblé (i.e. celui de l'emploi et du chômage) alors que le système ELEONORE (Rénié, Chanier, 1993) s'intéresse uniquement à la construction des interrogatives. En termes d'analyse et de feed-back, des critiques subsistent certainement mais les résultats restent bien plus précis et pertinents que pour l'analyse de textes libres. Le problème se situe davantage au niveau du ratio coût de développement/apport, d'autant plus que ces démarches semblent peu généralisables et exportables à d'autres types d'activités.

Ce tour d'horizon des diverses tentatives d'utilisation du TAL dans la détection/correction des fautes ne serait pas complet si nous ne citions pas les correcteurs orthographiques et grammaticaux. Contrairement aux systèmes évoqués, ces correcteurs n'ont pas une visée didactique, mais ils ont cependant donné lieu à différentes tentatives d'utilisation en classe de langue. Le constat semble relativement partagé (Cordier-Gauthier, Dion, 2003 ; Charnet, Panckhurst, 1998; Désilets, 1998) ; la relative qualité des analyses tant par le bruit et le silence générés que par les rétroactions inadaptées de ce type de logiciels permettent à

l'enseignant d'amorcer une réflexion avec les élèves autour des fautes détectées mais n'autorisent en aucun cas un travail en autonomie. En résumé, l'analyse automatique d'erreurs d'apprenants se heurte à deux problèmes majeurs. D'une part, des analyses TAL peu fiables car trop bruitées ou trop silencieuses et, d'autre part, comme le souligne également J. Rézeau (2001), du fait des coûts de recherche et de développement, ces projets restent souvent à l'état de prototypes voire de simples spécifications. Face à ces blocages, nous proposons une nouvelle approche du problème, la stratégie « moins-disante ». Les parties 3 et 4 de cet article seront consacrées à l'étude des premiers résultats d'ExoGen, un système fondé sur cette stratégie.

2 Stratégie « moins-disante »

La partie précédente de cet article met en évidence les faiblesses et les blocages actuels de l'analyse d'erreurs. Toutefois, nous tirons deux constats principaux de ces diverses expériences. Le premier concerne la bonne fiabilité des analyses TAL sous certaines conditions : cadre didactique contrôlé, lemmatisation, traitements des accords de courte portée, etc. Le deuxième constat prend appui sur les retours d'enseignants à l'usage de ces systèmes. L'exemple suivant en est relativement significatif :

« *Le Correcteur 101 est intéressant du fait qu'il amène l'élève à se questionner sur sa phrase et ses erreurs. Le logiciel n'offre que très rarement la réponse. Deux outils sont aussi offerts: un dictionnaire et une grammaire. Ils peuvent s'avérer très utiles pour dépanner. Le langage utilisé n'est pas conforme à la grammaire nouvelle, ce serait un atout pour permettre un lien à la grammaire étudiée en classe. Parfois, le logiciel ne mentionne pas certaines fautes.* »[1]

La détection des erreurs sans correction constitue un atout pédagogique en soi car il amène les apprenants à réfléchir sur leurs erreurs. La production de rétroactions (aides, explications, etc.) adaptées et didactiquement pertinentes facilitent cette réflexion et la recherche de solutions. Or, la qualité de ces rétroactions étant étroitement liée à la détection et à l'analyse des erreurs, ces deux phases doivent être contrôlées pour assurer une fiabilité et une précision maximales. La stratégie « moins-disante » que nous développons est un corollaire de ces constats. C'est une approche empirique qui se base sur les hypothèses suivantes :

- Il existe un ensemble de technologies TAL fiables (tokenisation, lemmatisation, étiquetage morpho-syntaxique…) pouvant fournir des éléments de détection et d'analyse des erreurs pertinents.
- L'analyse hors-contrôle du contexte de production n'est actuellement pas envisageable. Il est donc nécessaire de maîtriser, d'un point de vue final (i.e. didactique), les insuffisances du TAL. Dans notre cas, la connaissance des réponses attendues, qui permet des comparaisons, de lever des ambigüités, de cibler les analyses, etc., constitue une donnée du système.
- La fiabilité à 100% des analyses reste, et à notre avis restera, un objectif inaccessible. Pour éviter ce travers, nous préférons privilégier « l'aide à… » à l'automatisation des procédés. Effectivement, dans ce contexte, le TAL nous semble pouvoir améliorer sensiblement les différents niveaux de la chaîne d'ALAO : aide au choix d'exemples

[1] Critique d'une enseignante sur son utilisation du Correcteur 101 didactique en classe (http://c-rdi.qc.ca/produits/aff_fiche.asp?fiche=426, consulté le 31/01/2007). A noter que les autres critiques postées sur ce site vont globalement toutes dans ce sens.

(Antoniadis et al., 2007), aide à la génération d'activités (Antoniadis et al., 2005) ou encore l'aide à l'autocorrection (évaluation formative) qui constitue le sujet de cet article.

- En cas d'ambigüité, conserver la multiplicité des résultats pour une prise de décision au niveau didactique
- Développer une approche déclarative et modulaire des traitements pour permettre une évolution du système et de ses ressources.

La mise en œuvre de notre approche (cf. figure 1) est comparable à la reprise, par D. Anctil (2005), de « la stratégie de résolution de problèmes » développée par T. Andre (1986).

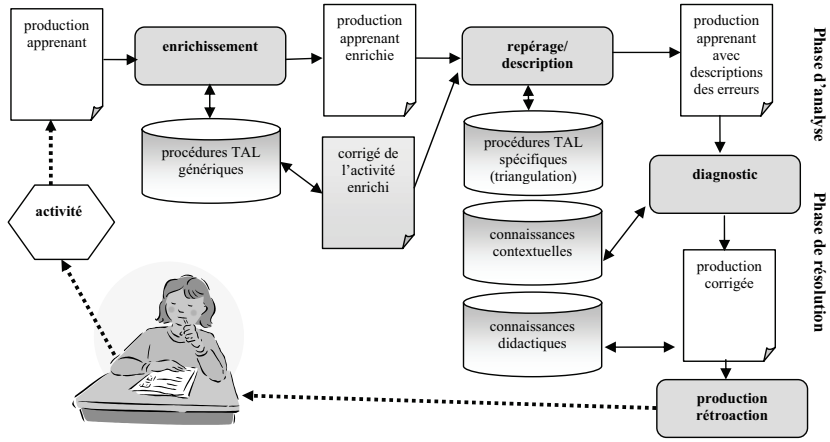


Figure 1 : Mise en œuvre de la stratégie « moins-disante »

Il s'agit de séparer la phase d'analyse du problème (i.e. l'erreur de l'apprenant pour notre système) de sa phase de résolution ; cette dernière passe par la production de rétroactions adaptées qui feront l'objet d'études ultérieures en collaboration avec des enseignants et des didacticiens. La phase de résolution permet, quant à elle, une distinction entre repérage, description et analyse de ce problème. Comme chez D. Anctil, nous avons réuni les étapes de repérage et de description du problème. En se basant sur un prétraitement[2] générique à la fois de la production de l'apprenant et du corrigé de l'activité, cette phase consiste en une analyse fine des différences entre réponses données (RD) et réponses attendues (RA). La désambiguïsation de ces analyses repose sur une triangulation entre RD, RA et le contexte.[3]

L'étape d'analyse du problème (appelée diagnostic) consiste, en fonction de connaissances sur le contexte de production (type d'activité, modèle apprenant, etc.) à rechercher les causes

[2] Il s'agit ici d'un enrichissement de la production de l'apprenant à l'aide de procédures de tokenisation, d'analyse morpho-syntaxique, etc.

[3] Par exemple, pour une activité portant sur l'accord du participe passé après l'auxiliaire avoir, si la forme correcte comporte des flexions (féminin ou pluriel) on peut en déduire que le verbe est transitif et qu'il existe un COD précédant le participe passé : ces inférences, étroitement liées à l'activité, permettent de désambiguïser les résultats d'une analyse classique (étiquetage) et d'appliquer ensuite des règles simplifiées pour le diagnostic.

potentielles de l'erreur et à sélectionner les informations pertinentes et fiables permettant un calcul de la rétroaction.

3 L'exemple du système ExoGen

Pour valider notre approche, nous avons développé un prototype, nommé ExoGen, qui permet la génération d'activités à partir d'un corpus de textes lemmatisés et étiquetés. Les activités proposées sont de type "lecture d'exemples" et "exercices lacunaires", et sont fondées sur l'extraction aléatoire de phrases contenant des patterns d'expressions régulières portant sur les formes, les lemmes, les catégories et les traits morphosyntaxiques (Kraif, 2006). Ces patterns permettent par exemple d'identifier des constructions telles qu'un participe passé dans un contexte de passé composé avec l'auxiliaire avoir, ou une flexion irrégulière pour le pluriel d'un nom, etc. A chaque génération d'activité, on peut obtenir un exercice lacunaire portant par exemple sur l'accord du participe passé avec l'auxiliaire avoir, les participes passés ayant été escamotés et remplacés par leur forme lemmatisée. Ce modèle de génération d'activité, très simple (analogue à celui du système Alfalex), se prête bien à l'analyse d'erreur précédemment décrite, basée sur la comparaison entre la réponse attendue RA et la réponse donnée RD.

Dans cette première implémentation, nous n'avons pas recouru au principe de triangulation, selon lequel l'analyse de la réponse attendue permettrait de désambiguïser à la fois le contexte et la réponse donnée, en vue d'appliquer une règle déclarative de diagnostic. Nous nous sommes contentés d'appliquer une heuristique simple lors de la phase d'analyse des différences, qui permet de privilégier, en cas d'ambiguïté, les analyses maximisant la similitude entre RA et RD (dans l'idée que cette similitude n'est pas fortuite). Qui plus est, nous n'utilisons pas les données issues de l'étiquetage et de la lemmatisation (obtenus avec Treetagger), afin de montrer jusqu'où cette heuristique permet de désambiguïser sans traitement préalable.

L'analyse et le diagnostic sont donc basés sur la seule comparaison de RA et de RD, indépendamment de tout contexte. Pour comparer ces deux formes nous ne disposons que d'une seule source d'information externe (en dehors des formes elles-mêmes) : les analyses possibles des formes fléchies, données par le dictionnaire de formes fléchies mis en ligne par l'ABU (http://abu.cnam.fr/). Chaque entrée de ce dictionnaire est une forme fléchie simple, à laquelle sont associés un lemme et les analyses possibles en terme de combinaison de traits morphosyntaxiques (nombre, genre, personne, temps, mode, etc.). La figure 2 donne un échantillon de quelques enregistrements de ce dictionnaire.

glace	glacer	Ver:IPre+SG+P1:IPre+SG+P3:SPre+SG+P1:SPre+SG+P3:ImPre+SG+P2
glacé	glacer	Ver:PPas+Mas+SG
glacent	glacer	Ver:IPre+PL+P3:SPre+PL+P3
glacera	glacer	Ver:IFut+SG+P3
glaceraient	glacer	Ver:CPre+PL+P3

Figure 2 : Un extrait du dictionnaire de formes fléchies

L'analyse repose sur une hiérarchisation des différences observées entre RA et RD : on traite en priorité les différences les plus légères et les plus superficielles, c'est-à-dire celles qui requièrent le moins d'inférence. A priori, ce sont aussi celles qui mènent aux diagnostics les plus sûrs :

1. **Différences graphiques**. Espacements, majuscules, variantes graphiques (p. ex. ligatures telle que oe et œ), etc. Ces différences, peuvent donner lieu à la validation de RD comme étant correcte, sauf si l'exercice porte explicitement sur ces aspects (usage des majuscules, réforme de l'orthographe, …).

2. **Différences orthographiques**. Si RD est absente du dictionnaire de formes fléchies, plusieurs cas peuvent être considérés :

a) RD ressemble à RA
  i/ RD est identique à RA si l'on néglige les signes diacritiques (accents, cédille)
  ii/ RD est une forme inconnue du dictionnaire
b) RD ne ressemble pas à RA
  i/ RD est une forme inconnue du dictionnaire, voisine de certaines formes connues
  ii/ RD est une forme inconnue du dictionnaire, et aucune forme voisine n'a été trouvée

La ressemblance peut être calculée par une fonction de Levenshtein, ou de recherche de la plus longue sous-chaîne commune (Kraif, 2001). Les formes voisines peuvent être trouvées grâce à une fonction de hachage dont les clés correspondraient à une écriture simplifiée (sans lettre double, sans accent, sans finales muettes, avec réduction des variantes graphiques, etc.). Chacun de ces cas peut donner lieu à un feed-back spécifique, en fonction du contexte didactique. Par exemple pour 2.a.i, on pourrait avoir "Accentuation incorrecte", pour 2.a.ii "Faute d'orthographe", pour 2.b.ii "Pensiez-vous à une de ces formes : [formes voisines]", etc.

3. **Différences morphosyntaxiques**. Si RD est connue dans le dictionnaire, on peut calculer ses lemmes potentiels, ainsi que les catégories et structures de traits afférentes. La comparaison peut alors porter sur chacun de ces aspects :

a) RD et RA correspondent à un même lemme
  i/ RD et RA correspondent à une même catégorie. Dans ce cas, seuls leurs traits les différencient.
  ii/ RD et RA correspondent à un même lemme, mais avec des catégories différentes. P.ex. RD="êtres" (être-Nom-PL), RA="êtes" (être-Ver-P2-PL).
b) RD et RA correspondent à des lemmes différents
  i/ RD et RA correspondent à une même catégorie avec les mêmes traits (p. ex. RA="trouverons" vs RD="penserons").
  ii/ RD et RA correspondent à une même catégorie avec des traits différents (p.ex. RA="conscience", RD="connaissances")
  iii/ RD et RA correspondent à des catégories différentes (p. ex. RD="mieux" (Adv) vs RA="préférable" (Adj)).

L'étude des cas est donc guidée par les similarités, le long d'un continuum allant de l'identité à la différence complète. Notre heuristique se fonde sur l'idée sous-jacente que les similarités sont rarement fortuites, tandis que les différences, elles, sont plus difficiles à systématiser. On peut en tirer une méthode de désambiguïsation lors de la comparaison des traits ou des catégories. Par exemple :

RA : *si j'avais su*	Catégorie : Ver	Traits : IImp+SG+P1	ou	IImp+SG+P2
RD : *si j'aurais su*	Catégorie : Ver	Traits : CPre+SG+P1	ou	CPre+SG+P2

Ici, on observe des différences au niveau du temps/mode ainsi que pour la personne : P1 ≠ P2, CPre ≠ IImp. On peut donc avoir 4 analyses différentes pour le couple (RA, RD). Grâce à l'heuristique de moindre différence, on ne compare que les analyses les plus proches (entre pointillés), ce qui permet de ne retenir qu'une seule différence de trait, entre le conditionnel

présent et l'imparfait : CPre ≠ IImp. Le feed-back correspondant pourrait être "Dans ce contexte, utilisez un imparfait plutôt qu'un conditionnel présent".

L'analyse pourrait se poursuivre sur le plan sémantique : lorsqu'on trouve deux lemmes distincts, de même catégorie, on peut évaluer leur proximité sémantique, par exemple en se basant sur une ressource du type dictionnaire de synonymes, thésaurus, réseau sémantique, etc. (cette fonctionnalité n'est pas encore implémentée dans notre prototype). Dans le cas d'unités polysémiques, l'heuristique de moindre différence permet encore de désambiguïser. Par exemple, si RA="pomme de pin", RD="pignon", et si le dictionnaire propose plusieurs acceptions pour "pignon" (/fruit/, /engrenage/) la comparaison RD/RA permettra de choisir l'acception la plus proche (i.e. le chemin le plus court à travers le graphe représentant les relations sémantiques).

4 Évaluation

Pour évaluer cette méthode simple d'analyse et de désambiguïsation des erreurs, il nous faudrait un corpus de réponses d'apprenants, obtenues dans le cadre d'activités de type QROC (questions à réponse ouverte courte), telles que des exercices lacunaires ou des quiz. Pour chaque réponse donnée, on pourrait ainsi appliquer l'analyse des erreurs en fonction de la réponse attendue. Bien que la constitution d'un corpus de ce genre soit prévue dans les développements ultérieurs d'ExoGen, nous ne disposons pas encore de telles données empiriques. Pour l'évaluation, nous avons donc utilisé une autre ressource, à savoir des exemples issus du corpus FRIDA (*FRench Interlanguage DAtabase*), constitué dans le cadre du projet Freetext (Granger et al., 2001). Il s'agit d'un corpus de rédaction d'apprenants de différents niveaux et de différentes langues maternelles pour lequel les erreurs ont été identifiées manuellement et balisées en fonction d'une typologie indiquant le domaine (morphologie, grammaire, lexique, etc.), la catégorie (agglutination, graphie, genre, etc.) et la catégorie grammaticale de la forme erronée. Pour chaque erreur identifiée, les annotateurs ont indiqué une correction. Ce corpus permet donc d'extraire des couples forme erronée / forme corrigée comparables aux couples RD / RA issus des QROC. On en tire des exemples du type :

> *(…) une seule monnaie (l'ECU) n'**adresse** pas bien au gouvernement anglais.*
> Forme erronée *: adresse,* Forme corrigée : *convient*

Cette évaluation peut comporter un biais, car le rapport RD/RA n'est pas identique à celui Réponse erronée/Réponse corrigée : dans le cadre d'un CROQ, la réponse attendue et son contexte préexistent à la réponse donnée, tandis qu'ici, les réponses corrigées sont données *a posteriori*, en fonction d'une erreur et d'un contexte issu d'une production libre. Mais nous pensons que ce biais est limité du point de vue de l'analyse des différences, car le même genre d'écart est observé, et la méthode d'analyse est confrontée au même type d'ambiguïtés (de lemme, de catégorie, de traits, de sens).

Nous avons utilisé un échantillon de 47 productions d'apprenants anglophones, de niveau variable. Ont été retenues toutes les erreurs impliquant deux formes simples (du fait des limitations de notre dictionnaire), hors ponctuation, pour un total de 318 cas d'erreurs. Pour chaque erreur nous avons appliqué l'analyse des différences, et obtenu des descriptions correspondant à 16 cas possibles, avec des précisions concernant les lemmes, catégories ou traits identifiés. On obtient par exemple les sorties suivantes :

Exemple d'erreur	Description (obtenue automatiquement)
(…) avant de **retourner** *(arriver)* en Angleterre.	Forme grammaticalement correcte (verbe infinitif), mais on attendait une autre forme.
et beaucoup d'**échafaide** *(échafaudages)*.	Orthographe erronée ou mot inconnu du dictionnaire.
Je dois me **dépécher** *(dépêcher)*.	Orthographe erronée : problème d'accent.
(…) sommes bien amusées et c'est **vrai** *(juste)* de dire que nous avons dansé assez bien	Forme grammaticalement correcte (adjectif ou adverbe ou nom masculin singulier), mais on attendait une autre forme
C'était **désespéré** *(désespérant)* mais c'était la seule chance (…)	S'il s'agit du verbe *désespérer* : cas 1 [masculin singulier] : On attend un participe présent et non un participe passé.
Pour moi **l'** *(cette)* image crée une ambiance délassante	Forme grammaticalement correcte sur le plan de la catégorie (déterminant), mais on attendait une autre forme avec d'autres traits.
le Premier ministre reste toujours un **britannique** *(Britannique)*	Exact, mais il faut une majuscule à l'initiale.

Tableau 3 : Exemples d'erreurs (corrigées entre parenthèses) et descriptions correspondantes

On constate que dans certains cas la désambiguïsation est partielle, ce qui n'empêche pas de donner une description pertinente. Pour une évaluation chiffrée des résultats, nous avons évalué manuellement la correction des affirmations liées aux différentes analyses. En outre, nous avons noté, pour tous les cas où les formes (erronées et corrigées) recelaient des ambiguïtés (analyses multiples), si la désambiguïsation est totale, partielle ou nulle.

	Tous les cas	Non ambigus	Complètement désambiguïsés	Partiellement désambiguïsés	Non désambiguïsés
Corrects	**312**	187	104	14	7
Incorrects	**6**	1	5	0	0
Précision	**0,981**	0,995	0,954	1	1

Tableau 4 : Evaluation de la correction des descriptions d'erreur

On constate que la précision est très satisfaisante. L'heuristique de désambiguïsation, opérante dans 1/3 des cas, aboutit très fréquemment à une désambiguïsation complète, avec un peu moins de 5% d'erreurs. Dans de nombreux cas, l'heuristique aboutit à une réduction spectaculaire des ambiguïtés :

*une seul monnaie (l'ECU) n' **adresse** (convient) pas bien au gouvernement anglais.*

Ici *adresse* peut correspondre à deux lemmes différents (*adresse* et *adresser*), deux catégories différentes (nom et verbe) et de nombreuses structures de traits (le dictionnaire en donne 6 : Nom:Fem+SG, Ver:IPre+SG+P1, IPre+SG+P3, SPre+SG+P1, SPre+SG+P3, ImPre+SG+P2). La comparaison avec *convient* permet de conserver la seule interprétation commune : verbe à l'indicatif présent, troisième personne du singulier. Quant aux analyses erronées, elles sont dues à deux phénomènes :

- Lacune du dictionnaire (2 cas) : dans l'exemple ci-dessous, le dictionnaire n'enregistre pas *futur* comme nom potentiel, mais seulement comme adjectif.

*le **futur** (avenir) de l'Angleterre -> "On attendait une autre forme, d'une autre catégorie grammaticale (Nom#Adj)."*

- Mauvaise désambiguïsation (4 cas) : dans l'exemple suivant, la forme corrigée est interprétée comme le déterminant *tous*, et non comme un pronom :

> *l'heure de se joindre et de parler **tout** (tous) d'une voix -> "S'il s'agit du **Det** tout on a : cas 1 [Masculin] : On attend un pluriel et non un singulier."*

Notons que même si la désambiguïsation est mauvaise, le feed-back donné à l'apprenant peut présenter une analyse comme hypothétique, ce qui évite d'affirmer une contre-vérité. Par ailleurs certaines ambiguïtés peuvent être réduites en sélectionnant les informations données à l'utilisateur. Prenons l'exemple suivant :

> *Soudain, nous avons **entendus** (entendu) un bruit -> "S'il s'agit du verbe entendre [participe passé masculin], on attend un singulier et non un pluriel ; S'il s'agit de l'adjectif entendu [masculin] on attend un singulier et non un pluriel ; s'il s'agit du nom entendu [masculin] on attend un singulier et non un pluriel."*

L'analyse donne un résultat ambigu (verbe, adjectif ou nom) mais l'analyse concernant les traits est toujours la même, et peut aboutir au feed-back suivant : "on attend un singulier et non un pluriel". Sur la base de la stratégie moins-disante, on pourra se contenter de cette information, incomplète mais fiable, et centrée sur l'erreur commise par l'apprenant.

5 Conclusion

Nous avons présenté un cadre général pour l'analyse des réponses d'apprenant, basé sur la comparaison entre réponse donnée et réponse attendue, en se limitant à des productions contrôlées, tant sur le plan de la forme qu'au niveau du contexte didactique. Ces limitations permettent selon nous de mettre en œuvre des techniques fiables dont les analyses n'outrepassent pas les capacités actuelles des systèmes de TAL. Cette approche, qualifiée de "moins disante" permet selon nous, lorsque le contexte didactique est suffisamment spécifié (de la définition d'une consigne à la mise en oeuvre de rétroactions adaptées), de diagnostiquer des fautes d'orthographe, des confusions, des problèmes d'accord, de conjugaison, etc. Dans cette optique, nous avons implémenté et évalué une méthode simple d'analyse des réponses, avec une heuristique de désambiguïsation tirant parti des écarts avec les réponses attendues. Les résultats sont encourageants, avec une précision supérieure à 98%.

L'étape suivante consistera à mettre en œuvre des règles de diagnostic, afin de déterminer les causes probables des erreurs (p.ex. un accord erroné d'un participe passé avec le sujet de l'auxiliaire avoir). Pour qu'un tel système soit généralisable et puisse s'adapter facilement à des activités variées, il est important de définir un langage simple et déclaratif pour la mise en œuvre des règles de diagnostic : c'est à notre avis une condition indispensable à l'appropriation de ces techniques par les pédagogues, qui seuls sont qualifiés pour interpréter les erreurs, déterminer leurs causes, et ensuite définir des rétroactions adaptées.

Pour y parvenir, nous prévoyons de développer des techniques de désambiguïsation plus fines, basées sur une triangulation entre RA/RD et leur contexte linguistique. Cette étape, dépassant les outils génériques tels que l'étiquetage et la lemmatisation, constituera un module de TAL relativement autonome, et généralisable à de nombreux types d'activités en ALAO.

Références

ANCTIL D. (2005). *Maîtrise du lexique chez les étudiants universitaires : typologie des problèmes lexicaux et analyse des stratégies de résolution de problèmes lexicaux*. Mémoire de M.A. Faculté des Sciences de l'Education. Université de Montréal (Québec).

ANDRE T. (1986). *Problem solving and education*. San Diego, CA: Academic Press.

ANTONIADIS G., ECHINARD, S., KRAIF O., LEBARBÉ T. & PONTON C. (2005). Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO. *ALSIC*, vol. 8, n° 1. pp. 65-79.

ANTONIADIS G., KRAIF O., PONTON C., ZAMPA V. (2007). Un outil exploratoire de corpus d'apprenants. Actes de *UNTELE'07*. Université de Compiègne. 29-31 mars 2007 (à paraître).

CHARNET C., PANCKHURST R. (1998). Le correcteur grammatical : un auxiliaire efficace pour l'enseignant ? Quelques éléments de réflexion. *ALSIC,* vol. 1, 2. pp. 103-114.

DÉSILETS M. (1998). Que penser de l'utilisation des logiciels correcteurs à l'école? *Vie pédagogique*. No 107. avril-mai 1998. 9-11.

GRANGER, S., VANDEVENTER, A., HAMEL, M.-J. (2001). Analyse des corpus d'apprenants pour l'ELAO basé sur le TAL. *Traitement automatique des langues*. *42* (2). 609-621.

KRAIF O. (2006) Extraction automatique de lexique bilingue : application pour la recherche d'exemples en lexicographie*, Journées du CRTT*, Université Lyon 2, Lyon.

KRAIF O. (2001) Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation, TAL 42 :3, ATALA, Paris, pp. 833-867.

L'HAIRE S. (2004). Vers un feed-back plus intelligent, les enseignements du projet Freetext. Actes de la *journée TALAL*. Grenoble. 1-12.

MEUNIER L. E. (2000). La typologie des intelligences humaine et artificielle : complexité pédagogique de l'enseignement des langues étrangères dans un environnement multimédia. *Apprendre une langue dans un environnement multimédia.* Les éditions Logiques : sous la direction de Duquette L. et Laurier M. Outremont (Québec). 211-253.

RÉNIÉ D., CHANIER T. (1993). La modélisation de l'acquisition, une étape dans la construction de systèmes d'EIAO des langues: le cas des interrogatives en français langue seconde. *Environnements Interactifs d'Apprentissage avec Ordinateur*. Tome 1. Eyrolles. 123-134.

RÉZEAU J. (2001). *Médiatisation et médiation pédagogique dans un environnement multimédia. Le cas de l'apprentissage de l'anglais en Histoire de l'art à l'université*. Thèse de troisième cycle. Université Bordeaux II, France.

SELVA, T., CHANIER, T. (2000). Génération automatique d'activités Lexicales dans le système ALEXIA. *Sciences et Techniques Educatives*, *(STE)*, 7 (2). 385-412.

WYATT D. H. (1987). Applying pedagogical principles to CALL courseware development. *Modern Media in Foreign Language Education*. Wm. Flint Smith (dir.). Lincolnwood, IL : National Textbook. 85-98.

Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré

Anna KUPŚĆ
Université Paris3 / LLF, UMR 7110
et Académie Polonaise des Sciences
akupsc@univ-paris3.fr

Résumé. Nous présentons une expérience d'extraction automatique des cadres de sous-catégorisation pour 1362 verbes français. Nous exploitons un corpus journalistique richement annoté de 15 000 phrases dont nous extrayons 12 510 occurrences verbales. Nous évaluons dans un premier temps l'extraction des cadres basée sur la fonction des arguments, ce qui nous fournit 39 cadres différents avec une moyenne de 1.54 cadres par lemme. Ensuite, nous adoptons une approche mixte (fonction et catégorie syntaxique) qui nous fournit dans un premier temps 925 cadres différents, avec une moyenne de 3.44 cadres par lemme. Plusieurs méthodes de factorisation, neutralisant en particulier les variantes de réalisation avec le passif ou les pronoms clitiques, sont ensuite appliquées et nous permettent d'aboutir à 235 cadres différents avec une moyenne de 1.94 cadres par verbe. Nous comparons brièvement nos résultats avec les travaux existants pour le français et pour l'anglais.

Abstract. We present our work on automatic extraction of subcategorisation frames for 1362 French verbs. We use a treebank of 15000 sentences from which we extract 12510 verb occurrences. We evaluate the results based on a functional representation of frames and we acquire 39 different frames, 1.54 per lemma on average. Then, we adopt a mixed representation (functions and categories), which leads to 925 different frames, 3.44 frames on average. We investigate several methods to reduce the ambiguity (e.g., neutralisation of passive forms or clitic arguments), which allows us to arrive at 235 frames, with 1.94 frames per lemma on average. We present a brief comparison with the existing work on French and English.

Mots-clés : français, corpus arboré, sous-catégorisation verbale, lexique-grammaire.

Keywords: French, treebank, verbal subcategorization, lexicon grammar.

1 Introduction

Cet article présente une expérience préliminaire d'extraction de sous-catégorisations verbales pour le français à partir d'un corpus journalistique richement annoté (le corpus arboré de Paris7).

Un lexique syntaxique est une ressource qui contient l'information sur le potentiel combinatoire d'un prédicat (ex., le verbe *dormir* régit un seul argument, le sujet), mais aussi sur le type de ses arguments (ex., l'adjectif *fier* se combine avec un syntagme prépositionnel en *de*). Ces

informations varient d'une langue à l'autre, elles sont donc essentielles pour l'apprentissage et l'acquisition des langues. Pour le traitement automatique des langues (TAL), les informations sur la structure prédicative sont importantes dans la plupart des applications. (Briscoe & Carroll, 1993) estiment qu'environ la moitié des erreurs des analyseurs syntaxiques repose sur des informations insuffisantes concernant la structure argumentale, tandis que (Carroll & Fang, 2004) montrent une amélioration significative de la performance d'un parseur enrichi avec un tel lexique. Elles jouent également un rôle essentiel pour la génération automatique (Danlos, 1985), la traduction automatique (Han *et al.*, 2000), ou l'extraction d'information, cf. (Surdeanu *et al.*, 2003).

Néanmoins, ce type d'informations est toujours difficilement disponible. Traditionnellement, de telles ressources ont été développées par des experts humains, par ex., (Procter, 1978; Hornby, 1989) (pour l'anglais) ou les lexiques-grammaires du LADL (Gross, 1975; Guillet & Leclère, 1992) et le Dictionnaire explicatif et Combinatoire (DECFC) de (Mel'cuk *et al.*, 1984 1988 1992 1999) (pour le français), ce qui garantit leur bonne qualité, mais elles ne sont pas directement adaptées au traitement automatique. Par contre, les ressources informatisées développées en vue des applications TAL utilisent des méthodes statistiques, par ex. : (Bourigault & Frérot, 2005; Chesley & Salmon-Alt, 2005), ou semi-automatique (Sagot *et al.*, 2006), (pour le français) ce qui rend les résultats moins fiables.

Dans cet article, nous nous basons sur le corpus arboré de Paris7, cf. (Abeillé *et al.*, 2003; Abeillé & Barrier, 2004), pour obtenir les cadres de sous-catégorisation verbales pour le français. à notre connaissance, ce corpus n'a pas encore été exploité pour l'extraction de telles ressources lexicales. Le corpus arboré de Paris7 est un ensemble de textes journalistiques du Journal Le Monde (1989-1993), annotés aux niveaux morphologique et syntaxique, pour les constituants majeurs mais aussi pour les fonctions grammaticales. L'étiquetage a été validé par des experts humains, ce qui fait du corpus une ressource précieuse pour des recherches linguistiques mais aussi pour le développement d'outils de TAL. Pour l'acquisition des cadres de sous-catégorisation verbale, nous nous sommes basés sur une sous-partie du corpus qui comprend des annotations fonctionnelles, soit environ 15 000 phrases (environ 300 000 mots).

L'objectif de ce travail est d'obtenir une liste de cadres de sous-catégorisation utilisables par différents types de grammaires électroniques, ainsi qu'un lexique informatisé, fiable et de haute qualité, qui pourra servir en particulier pour l'évaluation d'autres lexiques syntaxiques obtenus automatiquement. Il s'agit également d'estimer l'ambiguïté des cadres de sous-catégorisation verbale (combien de cadres par verbe ?) et de rechercher les méthodes pour la réduire. Ceci nous permettra de préparer une ressource bien adaptée pour différentes applications TAL.

2 état de l'art

Les travaux sur l'acquisition du lexique syntaxique à partir de treebanks sont relativement nombreux, et la plupart utilise des données moins riches que les notres. Pour l'anglais, le lexique syntaxique le plus important extrait à partir du corpus arboré (Penn-II Treebank) a été obtenu par (O'Donovan *et al.*, 2004), comme une ressource supplémentaire de l'induction des grammaires lexicalisées. La même technique a été adoptée pour d'autres langues comme l'allemand et le chinois, inter alia. (Sarkar & Zeman, 2000) présentent les résultats d'apprentissage automatique de cadres de sous-catégorisation à partir du corpus arboré du tchèque (Prague Dependency Treebank), tandis que (Marinov, 2004) applique les mêmes techniques sur un treebank bulgare

(BullTreebank). Tous ces lexiques sont obtenus à partir des annotations syntagmatiques, i.e., les fonctions grammaticales sont assignées automatiquement en utilisant des méthodes statistiques. Notre tâche est différente car nous bénéficions des annotations fonctionnelles déjà existantes dans le treebank, sans intermédiaire probabiliste.

Pour le français, les efforts récents pour construire des électroniques lexiques syntaxiques se sont basés sur les méthodes probabilistes, cf. (Bourigault & Frérot, 2005), (Chesley & Salmon-Alt, 2005), ou automatiques (Sagot *et al.*, 2006). Elles utilisent des corpus qui n'ont que des informations catégorielles ou bien des fonctions sont assignées automatiquement (par un parseur), ce qui pose problème pour distinguer arguments et ajouts. Un autre pôle est représenté par les travaux sur l'informatisation et l'actualisation du lexique–grammaire de LADL, effectué par (Gardent *et al.*, 2006). C'est une ressource tout à fait précieuse (plus de 8000 lemmes) mais qui n'a pas à notre connaissance été entièrement informatisée ni surtout validée sur corpus.[1] Un autre lexique syntaxique de la large couverture (plus de 3700 verbes), Proton[2] développé selon l'approche pronominale, n'est pas directement utilisable dans les application TAL.

3 Extraction des cadres de sous-catégorisation verbale

3.1 Choix d'informations à extraire

La représentation des cadres de sous-catégorisation se fait différemment selon différents approches : certains modèles théoriques, comme LFG (grammaire lexicale fonctionnelle) privilégient une notation basée sur les fonctions (1a), d'autres comme le LADL privilégient une notation basée sur les catégories (1b), d'autres enfin, comme en HPSG (grammaire syntagmatique guidée par les têtes), ont une approche mixte (1c) :

(1) *laver* :

a. <SUJ, OBJ>

b. N0 V N1

c. <SUJ :NP, OBJ :NP>

Les deux premières représentations ne sont pas complètes parce que les fonctions et les catégories peuvent avoir plusieurs réalisations (par exemple le sujet peut être nominal ou phrastique, tandis qu'un NP postverbal peut être objet direct ou attribut). Comme nous disposons d'un corpus annoté et pour les catégories et pour les fonctions, nous adoptons une approche mixte pour obtenir l'information plus riche. La liste des fonctions et des catégories dans le corpus est indiquée dans le tab. 1. Nous ignorons la fonction MOD qui correspond toujours à des éléments non sous-catégorisés (des modifieurs). Parmi les réalisations possibles des fonctions, nous ignorons les cas avec COORD, puisque la coordination double est très rare.

Pour les compléments prépositionnels (P-OBJ), nous marquons le type de la préposition régie par le verbe. Ceci nous permettra de normaliser les cadres par rapport aux formes passives et actives. Le noyau verbal, VN, contient le verbe principal mais aussi des auxiliaires, éléments négatifs, et les clitiques pronominaux. Selon une suggestion de (Abeillé & Barrier, 2004), nous considérons que le dernier V est la tête sémantique du VN. Les clitiques ont également une fonction indiquée (au niveau du VN) quand ils correspondent aux arguments du verbe.

[1] Le DECFC de Montréal est également en cours d'informatisation mais ne comprend que 514 vocables, qui ne sont pas tous des verbes (avec des informations sémantiques et non seulement syntaxiques).

[2] `http ://bach.arts.kuleuven.be/PA/proton.html`

SUJ	NP, VPinf, Ssub, COORD
OBJ	NP, AP, AdP, VPinf, COORD, Sint, Ssub
DE-OBJ	VPinf, PP, Ssub, COORD
A-OBJ	VPinf, PP, COORD
P-OBJ	PP, AdP, COORD, NP
ATO	Srel, PP, AP, NP, VPpart, COORD, VPinf, Ssub
ATS	NP, PP, AP, AdP, VPinf, Ssub, COORD, VPpart, Sint

FIG. 1 – Liste des catégories possibles pour chaque fonction sous-catégorisée. Fonctions : SUJ (sujet), OBJ (objet direct), DE-OBJ (objet indirect en *de*), A-OBJ (objet indirect en *à*), P-OBJ (complément avec une autre préposition), ATO (attribut de l'objet), ATS (attribut du sujet)

3.2 Description de l'expérience

L'extraction de cadres de sous-catégorisation verbale est plus difficile pour le français que pour l'anglais, d'une part à cause des alternances de variantes avec les pronoms clitiques, d'autre part à cause d'un ordre des mots plus libre (un SN (NP) postverbal peut être un sujet inversé par exemple). Nous avons extrait les lemmes des phrases principales du corpus arboré annoté pour les fonctions. Les fonctions sont traitées comme des attributs des syntagmes et non comme des relations entre la tête et les syntagmes. Elles sont notées soit sur les syntagmes de même niveau que le VN (pour les dépendants du verbe) soit sur le VN lui-même (pour les pronoms clitiques). Puisque le VN contient aussi les auxiliaires, nous traitons le dernier verbe dans VN comme le verbe principal, tandis que les auxiliaires sont stockés afin de normaliser les cadres par rapport aux formes passives et actives.

Comme point de départ, nous avons utilisé les cadres extraits directement du corpus, sans aucune modification, et ensuite nous avons fait plusieurs tests pour compacter les cadres.

D'abord, nous avons dégroupé les fonctions accumulées par les clitiques dans le VN. S'il y a plusieurs clitiques attachés au verbe (ex. : le sujet et l'objet dans *Il l'a vue*), leur fonctions sont groupées dans un seul tag (SUJ/OBJ). Il faut donc les séparer. Les cas où un clitique apparaît sans fonction sont normalement ceux qui correspondent à des réfléchis figés (comme pour *s'évanouir*). Nous les conservons comme tels dans nos cadres. Enfin, on peut avoir dans la même phrase un clitique et un argument nominal de même fonction. Ainsi dans une phrase comme : *Paul en mange-t-il beaucoup ?*, on a deux sujets (*Paul* et *il*) et deux objets (*en* et *beaucoup*). Il faut donc éliminer les duplicats des fonctions dans les cadres. Finalement, il y a des cadres qui n'ont pas de sujet spécifié. C'est le cas pour les formes verbales à l'impératif, et nous avons complété leurs cadres avec SUJ. Seules deux lemmes apparaissent toujours sans sujet : il s'agit de *voici* et *voilà* qui sont analysés comme des formes verbales à l'indicatif, et qui ont donc des cadres spécifiques sans sujet.

Nous avons normalisé les cadres par rapport aux formes du passif. On a utilisé une liste de 62 verbes qui sont conjugués avec *être* pour distinguer les formes du passé et du passif. Ainsi si le verbe apparaît dans le corpus avec l'auxiliaire *être* mais qu'il se conjugue avec *avoir*, son cadre est considéré comme passif et transformé en forme active. On ajoute OBJ, tandis que le complément d'agent (s'il est présent), i.e., P-OBJ introduit par la préposition *par* ou *de*, est supprimé (le SUJ est déjà présent dans les deux cadres passifs), et on change ATS en ATO s'il y a un attribut.

Un deuxième type de normalisation concerne les arguments clitiques. Comme nous extrayons les catégories des arguments, nous obtenons dans un premier temps, un cadre différent pour une occurrence avec sujet clitique et une occurrence avec sujet nominal alors que c'est la même sous-catégorisation. Nous avons donc regroupé les résultats.

Nous avons aussi commencé la factorisation des compléments optionnels. Par exemple, si un même verbe a deux cadres SUJ et SUJ OBJ, nous considérons que l'objet est optionnel. On peut donc lui assigner le cadre SUJ (OBJ).

Certaines difficultés viennent des choix d'annotation du corpus. Par exemple, les syntagmes adverbiaux sous-catégorisés ont une fonction syntaxique associée mais pas les adverbes seuls. Donc l'adverbe *bien* n'est pas reconnu comme le complément dans la phrase *Elle va bien*. Les annotations du corpus ne sont faites que pour les arguments qui appartiennent au cadre du verbe de même niveau. Donc on va rater des cas de dépendance à distance comme : *Que peut faire le gouvernement ?* (puisqu'on va extraire deux OBJ pour *peut* et aucun pour *faire*). Tels cas sont cependant assez rares.

3.3 Résultats

Dans cette expérience nous avons utilisé uniquement les verbes dans les phrases principales, soit 1362 verbe lemmes (12510 tokens). Nous comparons une approche uniquement fonctionnelle de la sous-catégorisation (comme (1a)) et une approche mixte qui tient compte également des catégories (comme (1c)).

3.3.1 Extraction de la Sous-catégorisation fonctionnelle

Nous avons normalisé les cadres par rapport au passif et nous n'avons pas utilisé les catégories. Nous ne tenons pas compte de l'ordre des mots, c'est-à-dire que nous considérons un seul cadre pour *Jean pense à Marie* et *à qui pense Jean* que nous notons 'A-OBJ, SUJ' (avec les fonctions en ordre alphabétique). Si l'on tient compte des clitiques réfléchis, qui peuvent être figés, on aboutit à 39 cadres différents avec une moyenne de 1.75 cadres par verbe. Le lemme avec le plus des cadres, 18, est le verbe *être*. Plus de la moitié des verbes (63.3% des lemmes, soit 862 lemmes différents) sont non ambigus et ont un seul cadre.

On peut réduire le nombre des cadres en éliminant le clitique réfléchi pour les verbes qui ont un OBJ ou un A-OBJ correspondant. Le nombre de cadres différents au total (39) comme le maximum de cadres par lemme (18 pour *être*) ne changent pas, mais on réduit la moyenne de cadres par lemme à 1.68. Avec ceci nous arrivons à 65.1% des verbes (888 lemmes) à un seul cadre.

Ensuite, si l'on factorise les arguments optionnels, par exemple les cadres SUJ et SUJ-OBJ, on obtient plus de cadres différents possibles et moins de cadres différents pour chaque verbe. Pour l'objet optionnel, nous avons un cadre de plus, c'est-à-dire 40 cadres en général, avec une moyenne de 1.54 cadres par lemme, tandis qu'il y a 4 verbes à 10 ou plus cadres, qui sont effectivement parmi les plus ambigus du français (*être*, *passer*, *avoir*, *rendre*). Les résultats sont présentés dans le tableau 2. Les cadres pour les 4 verbes avec le plus de cadres sont dans le tableau 3.

Les cadres les plus fréquents sont ceux des verbes à un complément, tout d'abord ceux à objet

	# cadres	moyenne	max. cadres	verbes à un cadre	
				%	#
avec réfléchi	39	1.75	18 (*être*)	63.3%	862
sans réfléchi	39	1.68	18 (*être*)	65.1%	888
SUJ (OBJ)	40	1.54	17 (*être*)	68.9%	939

FIG. 2 – Résultats pour les cadres fonctionnels

```
être (17):  (OBJ), SUJ| A-OBJ, ATS, OBJ, SUJ| A-OBJ, ATS, SUJ| A-OBJ,
  OBJ, SUJ| A-OBJ, SUJ| ATS, DE-OBJ, OBJ, SUJ| ATS, DE-OBJ, SUJ| ATS,
  DE-OBJ, SUJ, refl| ATS, OBJ, P-OBJ, SUJ| ATS, OBJ, SUJ| ATS, P-OBJ,
  SUJ| ATS, SUJ| ATS, SUJ, refl| DE-OBJ, OBJ, SUJ| DE-OBJ, SUJ| OBJ,
  P-OBJ, SUJ| P-OBJ, SUJ
avoir (11):  (OBJ), SUJ| A-OBJ, DE-OBJ, OBJ, SUJ| A-OBJ, OBJ, SUJ|
  A-OBJ, SUJ| ATO, OBJ, SUJ| ATS, OBJ, SUJ| DE-OBJ, OBJ, P-OBJ, SUJ|
  DE-OBJ, OBJ, SUJ| DE-OBJ, SUJ| OBJ, P-OBJ, SUJ| P-OBJ, SUJ
passer (10):  (OBJ), SUJ| A-OBJ, DE-OBJ, SUJ| A-OBJ, OBJ, SUJ| A-OBJ,
  SUJ| ATS, SUJ| DE-OBJ, SUJ| DE-OBJ, SUJ, refl| OBJ, P-OBJ, SUJ|
  P-OBJ, SUJ| P-OBJ, SUJ, refl
rendre (10):  A-OBJ, DE-OBJ, OBJ, SUJ| A-OBJ, OBJ, SUJ| A-OBJ, SUJ|
  ATO, DE-OBJ, OBJ, SUJ| ATO, OBJ, SUJ| ATS, P-OBJ, SUJ| ATS, SUJ|
  DE-OBJ, OBJ, SUJ| OBJ, SUJ| P-OBJ, SUJ, refl
```

FIG. 3 – Cadres fonctionnels pour les 4 verbes le plus ambigus : 10 cadres ou plus

direct (plus de la moitié des occurrences), puis ceux à sujet seul (le quart des lemmes), et ceux à objet indirect introduit par *à* ou *de*, et différents types de ditransitifs. Il y a relativement peu de verbes à attribut, mais ils sont très fréquemment utilisés. Dans le tableau 4, les cadres sont représentés avec les fonctions dans l'ordre alphabétique.

L'inconvénient de cette approche est que l'on a perdu des informations par rapport au corpus, en particulier on ne distingue pas les verbes qui prennent seulement un sujet nominal et ceux qui prennent un sujet nominal et phrastique. C'est pourquoi nous passons à une approche mixte.

3.3.2 Extraction de la Sous-catégorisation mixte

Nous avons adopté une représentation mixte, qui garde les fonctions et les catégories. Sans aucune factorisation, ni pour le passif ni pour les clitiques (pronominaux et réfléchis), on obtient 925 cadres différents, avec une moyenne de 3.44 cadres par verbe, et 49% des verbes (668 lemmes) qui n'ont qu'un seul cadre.

Après le dégroupage et l'élimination des duplicats des fonctions décrits dans la sec. 3.2, et avec la normalisation du passif, on réduit le nombre de cadres presque de moitié (on obtient 465 cadres différents), avec une moyenne de 2.78 cadres par verbe. Le nombre de verbes qui ont un seul cadre n'augmente pas beaucoup : 53% de verbes ne sont pas ambigus, soit 727 lemmes.

Nous procédons alors à la factorisation par rapport aux différentes réalisations clitiques (pour les sujets, les objets directs, les *de*-objets et les *à*-objets). Le taux d'ambiguïté baisse à 2.17 et nous obtenons 127 cadres de moins (337), tandis qu'un peu plus de 100 verbes ont un seul cadre

cadre	# types de verbes	occurrences
OBJ, SUJ	986 (72.4%)	6625 (52.9%)
SUJ	346 (25.4%)	1052 (8.4%)
A-OBJ, OBJ, SUJ	184 (13.5%)	423 (3.4%)
A-OBJ, SUJ	136 (9.9%)	423 (3.4%)
DE-OBJ, SUJ	125 (9.1%)	534 (4.3%)
OBJ, P-OBJ, SUJ	101 (7.4%)	165 (1.3%)
DE-OBJ, OBJ, SUJ	98 (7.2%)	196 (1.5%)
P-OBJ, SUJ	81 (5.9%)	215 (1.7%)
ATO, OBJ, SUJ	42 (3.1%)	1929 (15.4%)
SUJ, refl	36 (2.6%)	259 (2.1%)

FIG. 4 – Les 10 cadres de sous-catégorisation fonctionnelle les plus fréquents

	# cadres	moyenne	max. cadres	verbes à un cadre	
				%	#
données brutes	925	3.44	242 (*être*)	49%	668
normalisation passif	465	2.78	114 (*être*)	53.3%	727
normalisation clitiques	337	2.17	76 (*être*)	60.8%	829
normalisation réfléchi	337	2.11	76 (*être*)	62.5%	851
avec OBJ optionnel	338	2.00	75 (*être*)	64.4%	877
sans prépositions	235	1.94	62 (*être*)	64.4%	877

FIG. 5 – Résultats pour les cadres mixtes

être (62), avoir (23), rester (22), faire (17), passer (14),
trouver (14), estimer (13), sembler (13), rendre (13), devenir (11),
demander (11), aller (11), porter (11), déclarer (10), laisser (10)

FIG. 6 – Nombre de cadres mixtes pour 14 verbes les plus ambigus (10 cadres ou plus)

(829). La normalisation par rapport au clitique réfléchi diminue un peu l'ambiguïté (à 2.11 en moyenne) et augmente légèrement le nombre de verbes à un seul cadre (à 851). Le nombre de cadres ne change pas. Nous procédons enfin à la factorisation des objets optionnels, en ajoutant les cadres correspondants, ce qui nous amène à 338 cadres distincts, avec une moyenne de 2 cadres par verbe. Si, enfin, on neutralise la valeur lexicale des prépositions (différentes de *à* ou *de*), on obtient 235 cadres différents au total, et 1.94 en moyenne. Il reste 14 verbes avec plus de 10 cadres, avec un maximum de 62 cadres pour le verbe *être*, indiqués dans le tableau 6. Les résultats sont groupés dans le tableau 5.

Il est clair que l'approche mixte est plus précise mais aboutit à un grand éclatement par rapport à l'approche fonctionnelle : si l'on inclut les catégories, même avec les résultats les plus compactés, nous avons presque 6 fois plus des cadres ! Néanmoins, les taux d'ambiguïtés en moyenne et surtout les nombres de verbes avec un seul cadre sont relativement proches. Ceci nous donne l'espoir que l'approche mixte, plus riche en information, peut être adoptée dans les applications pratiques. Il nous reste certaines factorisations à effectuer : celles qui concerne l'optionalité des autres compléments, et celle qui concerne les attributs. En effet, on distingue les attributs selon leur catégorie, alors que pour *être*, par exemple, il s'agit du même cadre.

cadre	# types de verbes	occurrences
OBJ :NP, SUJ :NP	764 (55.6%)	3489 (27.8%)
SUJ :NP	159 (11.6%)	846 (6.7%)
A-OBJ :PP, OBJ :NP, SUJ :NP	103 (7.5%)	268 (2.1%)
OBJ :Ssub, SUJ :NP	92 (6.7%)	420 (3.3%)
DE-OBJ :PP, SUJ :NP	85 (6.2%)	308 (2.4%)
OBJ :VPinf, SUJ :NP	77 (5.6%)	1636 (13.7%)
P-OBJ :PP, SUJ :NP	73 (5.3%)	170 (1.3%)
OBJ :NP, P-OBJ :PP, SUJ :NP	68 (4.9%)	100 (0.8%)
A-OBJ :PP, SUJ :NP	68 (4.9%)	175 (1.3%)
SUJ :NP, refl :CL	36 (2.6%)	234 (1.9%)

FIG. 7 – Les 10 cadres de sous-catégorisation mixte les plus fréquents

Mais il n'est pas vrai que tous les verbes attributifs acceptent des attributs de n'importe quelle catégorie et il faut sans doute affiner les cadres (Lamiroy & Melis, 2005). On pourrait de même regrouper les compléments phrastiques et infinitifs pour les lemmes qui acceptent les deux.

Si l'on considère les cadres de sous-catégorisation les plus fréquents (tab. 7), on voit que comme dans l'approche précédente, c'est le cadre transitif direct qui arrive en tête. On voit aussi que le complément phrastique concerne plus de lemmes que le complément infinitif, mais beaucoup moins d'occurrences.

3.4 Discussion

Les approches précédentes ne disposant pas de la distinction entre ajouts et arguments dans le corpus de départ, adoptent une approche statistique qui les conduit à ignorer les lemmes à fréquence basse (moins de 5 occurrences). Puisque nous disposons de ces informations dans le corpus, ceci nous permet de considérer aussi les cadres plus rares.

3.4.1 Comparaison avec les travaux sur l'anglais

Pour l'anglais, le taux d'ambiguïté de cadres est rarement mentionné. (Manning, 1993) rapporte la moyenne de 1.43 cadres par verbe pour le lexique de 1856 lemmes, ce qui est comparable avec les chiffres que nous obtenons : un lexique de 1362 lemmes et 1.54 cadres en moyenne (cadres fonctionnels). La différence principale réside non seulement dans la méthode adoptée (l'approche statistique) mais aussi dans le fait qu'il utilise 19 cadres présupposés, tandis que nous les acquérons à partir des annotations dans le corpus. (O'Donovan *et al.*, 2004) adoptent, comme nous, une approche mixte pour la représentation des cadres et obtiennent un lexique de 4362 lemmes, avec environ 4 cadres en moyenne, 38 types de cadres basés uniquement sur les fonctions et 577 cadres acquis si les différents types de prépositions et particules sont inclus. Les auteurs sont quand même obligés à adopter une méthode automatique pour obtenir les annotations fonctionnelles car ces informations ne sont pas présentes dans le corpus.

3.4.2 Comparaison avec les travaux sur le français

Les tables du LADL comportent 38 cadres principaux pour les verbes simples. Ces cadres sont basés sur la catégorie des arguments et non sur les fonctions, et ils tiennent compte de la valeur lexicale de certaines prépositions. Les tables distinguent ainsi parmi les compléments, les cadres où seul un complément infinitif est autorisé (table 1) ou ceux avec un complément nominal introduit par la préposition *à* (table 33). Nos résultats, obtenus par la méthode fonctionnelle, sont presque de même taille (39 cadres différents) mais ils sont en fait différents. Nous avons par exemple des cadres pour les emplois attributifs (cadres avec attribut du sujet ou attribut de l'objet) qui ne figurent pas dans les tables du LADL. D'autre part, nous avons certains emplois figés (cf. verbes avec clitique figé) ce qui crée des cadres supplémentaires. Si l'on compare les tables avec nos résultats obtenus par la méthode mixte, nos chiffres sont très supérieurs. Il y a deux raisons à ceci : d'une part nous avons dégagé des cadres supplémentaires par rapport aux tables LADL (pour les attributs, pour les clitiques figés ou pour les verbes sans sujet) et d'autre part il nous reste encore à faire certains regroupements, par exemple pour les catégories des attributs, ou les emplois à complément infinitif ou phrastique.

(Candito, 1999) et (Abeillé, 2002) décrivent les familles d'arbres de la grammaire FTAG. Il s'agit pourtant de cadres abstraits qui ne sont pas couplés à un lexique de grande taille. Dans la grammaire FTAG sont distinguées 45 familles à tête verbale, dont 15 à arguments nominaux et 24 à arguments phrastiques et 6 à complément adverbial. Cette grammaire inclut, comme ici, un cadre pour les formes verbales sans sujet, *voici* et *voilà*, quelques cadres pour les verbes à clitique figé (comme *s'évanouir* ou *s'appeler* N). Mais il est clair que nous extrayons des cadres supplémentaires.

4 Conclusion

Les résultats préliminaires d'extraction des cadres de sous-catégorisation verbale à partir de treebank français présenté dans cet article sont encourageants. Nous avons réussi à considérablement réduire l'ambiguïté de la représentation mixte avec les différentes techniques de factorisations. Ce résultat nous laisse espérer qu'il est donc possible d'incorporer les informations plus riches dans les applications pratiques.

Nous planifions plusieurs extensions de ce travail. Pour obtenir un lexique plus important, nous allons prendre toutes les occurrences des verbes (environ 2200 verbes, soit 15% de 15000 verbes en usage, selon (Gross, 1975; Guillet & Leclère, 1992)) et non seulement les verbes dans les phrases principales. On prévoit aussi l'extraction d'autres catégories prédicatives (cadres de sous-catégorisation des adjectifs ou des noms). Nous pensons incorporer ce lexique dans les applications TAL. Les cadres obtenus peuvent être facilement traduits dans un modèle et un format utilisables par un analyseur syntaxique (par exemple XLFG de L. Clément basé sur LFG), dans une autre application ou selon le format spécifié par le projet LexSynt.[3] Nous proposons aussi de confronter nos résultats à d'autres corpus (par exemple, TLFI ou Frantext) pour valider le lexique et le comparer avec d'autres travaux. Ceci nous permettra aussi enrichir le treebank de départ en ajoutant automatiquement le cadre de sous-catégorisation à chaque occurrence verbale (quand on en a extrait un seul) ou en en ajoutant plusieurs (avec un choix à valider par un annotateur humain) si l'on en a extrait plusieurs.

[3] `http ://lexsynt.inria.fr/index.php`

Références

ABEILLÉ A. (2002). *Une grammaire électronique du français*. CNRS Editions.

ABEILLÉ A. & BARRIER N. (2004). Enriching a French treebank. In *Proceedings of the LREC04 Conference*, Lisbonne.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In *Treebanks*. Kluwer.

BOURIGAULT D. & FRÉROT C. (2005). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*.

BRISCOE T. & CARROLL J. (1993). Generalised probabilistic LR parsing for unification-based grammars. *Computational linguistics*.

CANDITO M.-H. (1999). *Répresentation modulaire et parametrable de grammaires électroniques lexicalisées. Application au français et à l'italien.* PhD thesis, Université Paris7.

CARROLL J. & FANG A. (2004). The automatique acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Conference on Natural Language Processing*, Sanya City, China.

CHESLEY P. & SALMON-ALT S. (2005). Le filtrage probabiliste dans l'extraction automatique de cadres de sous-catgorisation. In *Journé ATALA sur l'interface lexique-grammaire*, Paris.

DANLOS L. (1985). *La generation automatique de textes*. Masson.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2006). Extraction d'information de sous-catégorisation à partir du lexique-grammaire de Maurice Gross. In *TALN 2006*.

GROSS M. (1975). *Méthodes en syntaxe*. Hermann.

GUILLET A. & LECLÈRE C. (1992). *La structure des phrases simples en français*. Genève : Droz.

HAN C., YOON J., KIM N. & PALMER M. (2000). *A Feature-Based Lexicalized Tree Adjoining Grammar for Korean*. Rapport interne, IRCS.

HORNBY A. S. (1989). *Oxford Advanced Learner's Dictionary of Current English*. Oxford : Oxford University Press, 4th edition.

LAMIROY B. & MELIS L. (2005). Les copules ressemblent-elles aux auxiliaires ? In SHYLDKROT, H. BAT-ZEEV & N. L. QUERLER, Eds., *Les Périphrases Verbales*, p. 145–170.

MANNING C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31th Meeting of the ACL*, p. 235–242, Columbus, Ohio.

MARINOV S. (2004). Automatic extraction of subcategorization frames for Bulgarian. In P. EGRÉ & L. A. I ALEMANY, Eds., *Proceedings of the Ninth ESSLLI Student Session*.

MEL'CUK I., ARBATCHEWSKY-JUMARIE N. & CLAS A. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques, vol. I, II, III, IV*. Les Presses de l'Université de Montréal.

O'DONOVAN R., BURKE M., CAHILL A., VAN GENABITH J. & WAY A. (2004). Large-scale induction and Evaluation of Lexical Resources from the Penn-II Treebank. In *Proceedings of the 42nd Conference of the Association for Computational Linguistics*, p. 367–374, Barcelona, Spain.

P. PROCTER, Ed. (1978). *Longman Dictionary of Contemporary English*. Burnt Mill, Harlow : Longman.

SAGOT B., CLÉMENT L., DE LA CLERGERIE E. V. & BOULLIER P. (2006). The lefff 2 syntactic lexicon for french : architecture, acquisition, use. In *Actes de LREC 06, Gênes, Italie*.

SARKAR A. & ZEMAN D. (2000). Automatic extraction of subcategorization frames for Czech. In *Proceedings of Colling 2000*.

SURDEANU M., HARABAGIU S., WILLIAMS J. & AARSETH P. (2003). Using predicate-argument structures for information extraction.

Vers une formalisation des décompositions sémantiques dans la Grammaire d'Unification Sens-Texte

François LAREAU
Lattice - U. Paris 7, UFRL, Case 7003, 2 pl. Jussieu, 75251 Paris cedex 5
OLST - U. de Montréal, Ling., CP 6128 succ C.-V., Montréal QC, H3C 3J7
francois.lareau@umontreal.ca

Résumé. Nous proposons une formalisation de la décomposition du sens dans le cadre de la Grammaire d'Unification Sens-Texte. Cette formalisation vise une meilleure intégration des décompositions sémantiques dans un modèle global de la langue. Elle repose sur un jeu de saturation de polarités qui permet de contrôler la construction des représentations décomposées ainsi que leur mise en correspondance avec des arbres syntaxiques qui les expriment. Le formalisme proposé est illustré ici dans une perspective de synthèse, mais il s'applique également en analyse.

Abstract. We propose a formal representation of meaning decomposition in the framework of the Meaning-Text Unification Grammar. The proposed technique aims at offering a better integration of such semantic decompositions into a global model of the language. It relies on the saturation of polarities to control the construction of decomposed respresentations as well as their mapping to the syntactic trees that express them. The proposed formalism is discussed from the viewpoint of generation, but it applies to analysis as well.

Mots-clés : Grammaire d'Unification Sens-Texte, Théorie Sens-Texte, sémantique, représentation du sens, paraphrasage.

Keywords: Meaning-Text Unification Grammar, Meaning-Text Theory, semantics, representation of meaning, paraphrasing.

1 Introduction

Dans le cadre de la théorie Sens-Texte (Mel'čuk, 1997; Kahane, 2001) (ci-après, TST), le sens d'un énoncé est représenté par un réseau de sémantèmes. Ces sémantèmes peuvent être décomposés (c'est-à-dire remplacés par leur définition) mais ce sont en général des représentations réduites (c'est-à-dire dont les sémantèmes ne sont pas décomposés) qui sont utilisées pour représenter le sens d'un énoncé. La principale raison pour laquelle on ne décompose pas les représentations sémantiques est qu'on ne saurait pas où s'arrêter. En effet, jusqu'à quel point faut-il décomposer ? Il n'y a à cette question que deux réponses possibles si on veut éviter l'arbitraire : soit on décompose jusqu'aux sens primitifs, ce que fait notamment (Wierzbicka, 1996), soit on ne décompose pas du tout, ce qui est la position de la TST (Mel'čuk, 1988; Mel'čuk, 1989). Ainsi, bien qu'ils soient des paraphrases, les trois énoncés suivants n'ont pas la même représentation sémantique, comme le montre la Figure 1.

(1) *a. Une Torontoise rousse.*

b. Une rousse qui habite Toronto.

c. Une femme rousse qui habite Toronto.

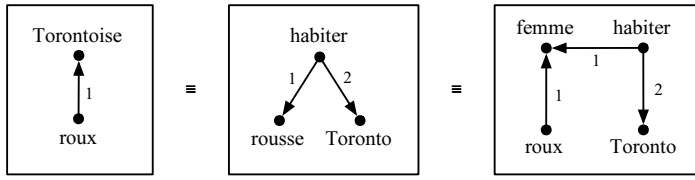


FIG. 1 – Trois paraphrases qui n'ont pas la même représentation sémantique

Ces trois énoncés expriment exactement les mêmes sens, soit ceux visibles à la Figure 1c, mais ils diffèrent par la façon dont ils les regroupent. Dans le premier les sens ⟨femme⟩, ⟨habiter⟩ et ⟨Toronto⟩ sont exprimés par un même nom, TORONTOISE$_{(N)}$ (*cf.* Figure 1a), alors que dans le second ce sont ⟨femme⟩ et ⟨roux$_{(Adj)}$⟩ qui sont regroupés et exprimés par le nom ROUSSE$_{(N)}$ (*cf.* Figure 1b), tandis que dans le troisième énoncé à chaque sémantème correspond un lexème distinct (*cf.* Figure 1c). Bien entendu, on peut démontrer que leurs représentations sémantiques respectives sont strictement équivalentes, tout simplement par le biais de ce que (Milićević, sous presse) appelle des règles d'expansion / réduction ordinaires, c'est-à-dire des règles qui mettent en équivalence les sémantèmes et leur définition. La Figure 2 montre un exemple d'une telle règle.

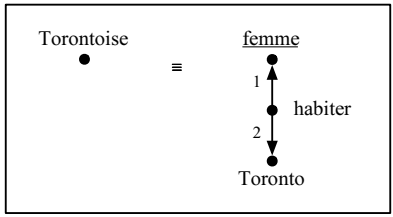


FIG. 2 – Une règle d'équivalence sémantique de la TST

Toutefois, ces règles soulèvent au moins trois questions quant à leur formalisation :

1. Où se situent les règles d'équivalence sémantique par rapport aux différents modules d'un modèle linguistique ?
2. Comment peuvent-elles interagir avec les autres règles du modèle ?
3. Comment peuvent-elles être formalisées ?

Nous tenterons d'y répondre en nous situant dans le cadre de la Grammaire d'Unification Sens-Texte, dont nous proposons d'abord un bref rappel. Nous suggérons une solution qui ne fait appel à aucun mécanisme extérieur et qui ne nécessite aucun changement au formalisme. Nous

verrons également brièvement comment notre solution peut servir à typer les actants sémantiques, si désiré.

2 La Grammaire d'Unification Sens-Texte

La Grammaire d'Unification Sens-Texte (ci-après, GUST) a d'abord été proposée par (Kahane, 2002), puis formalisée sous forme de grammaire d'unification polarisée (Kahane, 2004) par (Kahane & Lareau, 2005). Puisqu'elle est basée sur la TST, GUST considère plusieurs niveaux de représentation (sémantique, syntaxique, morphotopologique et phonologique). À chaque niveau correspond une grammaire de bonne formation qui décrit les structures possibles (ces grammaires sont appelées respectivement $\mathcal{G}_{\text{sém}}$, $\mathcal{G}_{\text{synt}}$, $\mathcal{G}_{\text{morph}}$ et $\mathcal{G}_{\text{phon}}$). Les niveaux sont ordonnés et la correspondance entre deux structures de niveaux adjacents est assurée par des grammaires d'interface ($\mathcal{I}_{\text{sém-synt}}$, $\mathcal{I}_{\text{synt-morph}}$ et $\mathcal{I}_{\text{morph-phon}}$) qui mettent en correspondance des fragments élémentaires de structures.

Tous les nœuds, arcs et autres objets des grammaires de bonne formation ou d'interface sont associés à un certain nombre de polarités. Contrairement à (Kahane & Lareau, 2005), nous utilisons un système à deux valeurs de polarités (plutôt que trois), soit □ et ■. Parmi ces deux valeurs, la noire est dite neutre, c'est-à-dire qu'une structure dont tous les objets ne portent que des polarités de valeur noire est considérée comme saturée, alors qu'une polarité blanche déclenche obligatoirement l'unification de son porteur avec d'autres objets afin d'être neutralisée. L'unification de deux objets entraîne en effet le calcul d'une nouvelle valeur pour chacune des polarités qu'ils portent de la façon suivante : □ · □ = □ et □ · ■ = ■ · □ = ■ (l'unification de deux objets noirs n'est pas permise). C'est ainsi que, par unification, les structures se saturent. C'est en fait la neutralisation des polarités qui contrôle tout le processus de synthèse ou d'analyse linguistique. Cela s'opère notamment par les polarités d'interface, c'est-à-dire des polarités non neutres qui sont ajoutées aux objets d'une grammaire dans le but de déclencher le module suivant dans le modèle linguistique. Par exemple, les objets de $\mathcal{G}_{\text{sém}}$, en plus de porter une polarité $p_{sém}$, qui contrôle la saturation des structures sémantiques, portent tous une polarité $p_{sém\text{-}synt}$ blanche. Les règles de $\mathcal{I}_{\text{sém-synt}}$ étant les seules à contenir des objets de polarité $p_{sém\text{-}synt}$ noire, ce module sera automatiquement appliqué, de façon à neutraliser les structures produites par $\mathcal{G}_{\text{sém}}$. Nous ne pouvons pas illustrer en détail ce mécanisme dans le présent article, aussi renvoyons-nous le lecteur à (Kahane & Lareau, 2005). Nous verrons qu'il est au cœur de la formalisation que nous proposons ici pour les règles d'équivalence sémantique.

3 La place des règles d'équivalence sémantique dans GUST

Dans le cadre de la TST, les règles d'équivalence sémantique ne font partie d'aucun des modules de transition qui constituent un modèle Sens-Texte, puisqu'elles encodent des phénomènes orthogonaux aux processus de synthèse et d'analyse simulés par les règles de correspondance. Pour les mêmes raisons, elles ne peuvent pas constituer une grammaire d'interface dans GUST, puisqu'elles ne mettent pas en relation des objets de niveaux adjacents. En fait, ces règles ne peuvent même pas constituer un module distinct, puisqu'on ne sait pas a priori jusqu'à quel niveau on doit décomposer (ou réduire) un sémantème (ou une configuration de sémantèmes). Il

faudrait donc que ce module puisse s'appliquer à son propre résultat de façon récursive, ce qui est difficilement compatible avec le système de polarités utilisé dans GUST. Ainsi, nous proposons plutôt d'intégrer les règles d'expansion / réduction à la grammaire de bonne formation sémantique de GUST.

4 Une représentation formelle du sens en GUST

Afin d'intégrer les règles d'expansion / réduction à $\mathcal{G}_{\text{sém}}$, nous devons légèrement modifier la caractérisation de celle-ci donnée par (Kahane & Lareau, 2005), sans pour autant en modifier le formalisme. Les nœuds sémantiques seront ici conçus comme des ensembles qui peuvent contenir d'autres nœuds. Les nœuds inclus à l'intérieur d'un autre nœud sont reliés par des arcs et forment de petites structures sémantiques qui représentent les constituants du sémantème ainsi décomposé. Les nœuds qui décomposent un autre nœud sont reliés à ce dernier par la fonction *réduction*, qui associe à un nœud l'ensemble dont il fait partie (c'est-à-dire un autre nœud). Parmi les nœuds qui forment la décomposition d'un sémantème, un seul est identifié comme le sens principal. Il s'agit du genre prochain et il est identifié par la fonction *générique*, qui prend comme argument un sémantème (un nœud) et retourne son genre prochain (un autre nœud).[1] Graphiquement, nous représentons les décompositions à l'intérieur d'une bulle en pointillés liée au nœud réduit, dans laquelle le genre prochain est souligné. Par souci de lisibilité, nous réduisons la taille des caractères et des traits à l'intérieur de ces bulles. La Figure 3 encode la règle d'équivalence ⁽Torontoise⁾ ≡ ⁽femme qui habite Toronto⁾, tant sous forme explicite[2] que graphique.[3]

{ 1:nœud$_{\text{sém}}$, 2: nœud$_{\text{sém}}$, 3: nœud$_{\text{sém}}$,
4: nœud$_{\text{sém}}$, 5: arc$_{\text{sém}}$, 6: arc$_{\text{sém}}$ },
{ *source*(5) = 2, *cible*(5) = 3,
source(6) = 2, *cible*(6) = 4,
étiquette(1) = "Torontoise",
étiquette(2) = "habiter", *étiquette*(3) = "femme",
étiquette(4) = "Toronto", *étiquette*(5) = "1",
étiquette(6) = "2",
$p_{sém}(1) = \text{N}, p_{sém}(2) = \text{B}, p_{sém}(3) = \text{B},$
$p_{sém}(4) = \text{B}, p_{sém}(5) = \text{B}, p_{sém}(6) = \text{B},$
$p_{sém\text{-}synt}(1) = \text{B}, p_{sém\text{-}synt}(2) = \text{N}, p_{sém\text{-}synt}(3) = \text{N},$
$p_{sém\text{-}synt}(4) = \text{N}, p_{sém\text{-}synt}(5) = \text{N}, p_{sém\text{-}synt}(6) = \text{N},$
générique(1) = 3, *réduction*(2) = 1,
réduction(3) = 1, *réduction*(4) = 1 }

≡

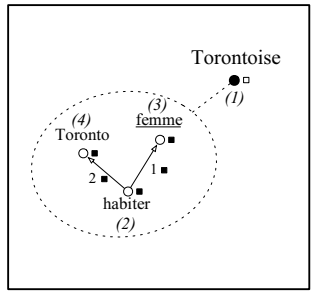


FIG. 3 – La décomposition sémantique de ⁽Torontoise⁾

[1] Il n'est pas évident que l'on puisse toujours identifier un seul sémantème comme genre prochain pour tous les sens de la langue. Nous ne savons pas pour l'instant comment encoder les cas où le genre prochain est une configuration de sémantèmes ou où il y a plusieurs genres prochains.

[2] Les valeurs « N » et « B » des fonctions de polarités renvoient respectivement aux polarités ■ et □.

[3] Toutes les figures qui suivent montrent deux polarités pour chaque objet : $p_{sém}$ (qui contrôle la saturation des structures sémantiques), indiquée par la couleur de l'objet, et $p_{sém\text{-}synt}$ (qui assure l'interface avec $\mathcal{I}_{\text{sém-synt}}$), indiquée par un petit carré à côté de l'objet. Par ailleurs, dans la Figure 3, nous ajoutons aux nœuds de la version graphique (entre parenthèses et en italique) les numéros d'identification utilisés dans la version textuelle afin que le lecteur puisse plus facilement mettre en relation les deux formes de représentation.

Les sémantèmes ⟨rousse$_{(N)}$⟩ et ⟨femme⟩ peuvent se décomposer de la même façon, comme le montre la Figure 4. On remarquera que les nœuds à l'intérieur des décompositions portent toujours une polarité $p_{sém}$ blanche. Ils forcent ainsi l'application d'autres règles de $\mathcal{G}_{\text{sém}}$ pour vérifier leur bonne formation. Par exemple, le nœud ⟨femme⟩ apparaissant dans la décomposition de ⟨rousse$_{(N)}$⟩ forcera l'application de la règle qui en vérifie la bonne formation (à droite dans la Figure 4), introduisant par le fait même la décomposition de ce sémantème.

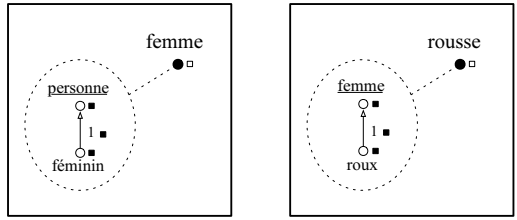


FIG. 4 – La décomposition sémantique de ⟨femme⟩ et de ⟨rousse$_{(N)}$⟩

On remarquera également que les nœuds à l'intérieur des décompositions sont toujours saturés en $p_{sém\text{-}synt}$. Cela a pour but d'éviter de réaliser à la fois un sens et sa définition dans le même énoncé. Lorsqu'un fragment de la structure sémantique a été réduit, tous ses éléments deviennent invisibles pour l'interface $\mathcal{I}_{\text{sém-synt}}$, qui ne peut plus voir que le sémantème réduit.

Ainsi, toutes les règles de $\mathcal{G}_{\text{sém}}$ qui construisent un nœud devront également en fournir la décomposition. Seuls les primitifs sémantiques[4] font exception, puisqu'ils ne peuvent pas être décomposés. Pour les besoins de notre exposé, nous allons considérer que ⟨(de sexe) féminin⟩, ⟨(aux cheveux) roux$_{(Adj)}$⟩, ⟨habiter⟩, ⟨Toronto⟩ et ⟨personne⟩ sont des sens indécomposables.

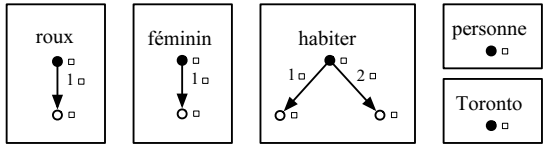


FIG. 5 – Les sens primitifs dans la grammaire sémantique

Voyons maintenant comment ces règles s'appliquent en synthèse. D'abord, les sens « primitifs » décrits à la Figure 5 nous permettent de construire la représentation de la Figure 6.

[4]Nous partons du postulat, hérité de la TST, que toute langue comporte un certain nombre de sens indécomposables, dits « primitifs » (qui ne sont pas forcément universels). Notre formalisme ne peut pas fonctionner si on admet la circularité dans les descriptions sémantiques. Cette propriété est d'ailleurs souhaitable à nos yeux puisqu'elle permet de vérifier la non-circularité de la grammaire sémantique.

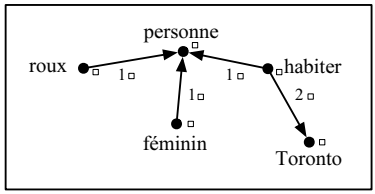


FIG. 6 – Une représentation sémantique décomposée à réaliser

Il est parfaitement possible d'utiliser cette représentation sémantique telle quelle et de la mettre en correspondance avec un arbre syntaxique sans la réduire aucunement. On obtient alors l'énoncé (2).

(2) *Une personne rousse de sexe féminin qui habite Toronto.*

On peut également opérer une première réduction en appliquant la définition de ⟨femme⟩ (voir la Figure 4), ce qui donne la structure de la Figure 7.

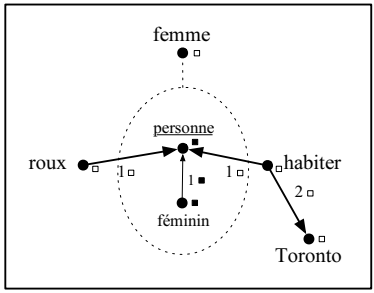


FIG. 7 – La réduction de ⟨personne (de sexe) féminin⟩

Cette représentation ne peut toutefois pas être mise en correspondance avec un arbre syntaxique, puisque du point de vue de l'inteface sémantique-syntaxe elle n'est pas connexe (⟨personne⟩ étant saturé en $p_{sém\text{-}synt}$, il n'est plus visible pour l'interface). La structure qui en résulterait ne serait pas un arbre et serait donc rejetée (Kahane & Lareau, 2005). Pour rendre cette structure connexe, il faut en quelque sorte « déplacer » les relations sémantiques qui pointaient vers ⟨personne⟩ pour qu'elles portent maintenant sur ⟨femme⟩. Nous devons donc introduire une règle de propagation qui recopie les relations sémantiques pointant vers un nœud pour qu'elles pointent vers le sémantème réduit dont ce nœud est le genre prochain[5] (voir Figure 8). Une fois cette règle appliquée, on obtient la structure de la Figure 9.

[5]Cette règle ne concerne que le genre prochain. En effet, si une relation pointe vers un autre sémantème, la réduction n'est pas possible. Par exemple, ⟨femme très rousse⟩ ne peut pas être réduit sans perte d'information [**une très rousse*$_{(N)}$], alors que ⟨grande femme rousse⟩ ne pose aucun problème [*une grande rousse*$_{(N)}$].

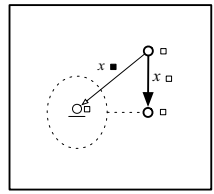


FIG. 8 – Une règle de propagation du gouverneur

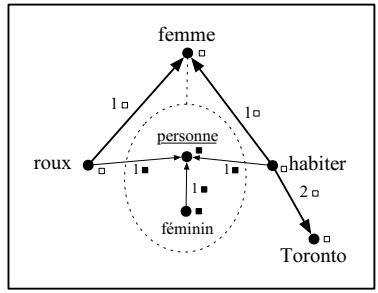


FIG. 9 – La réduction de ⟨personne (de sexe) féminin⟩ avec propagation des gouverneurs

Les relations pointant vers ⟨personne⟩ ont été neutralisées du point de vue de l'interface sémantique-syntaxe par la saturation de leur polarité $p_{sém\text{-}synt}$. La structure résultante peut tout de suite être mise en correspondance avec un arbre syntaxique, ce qui donne l'énoncé (3).

(3) *Une femme rousse qui habite Toronto.*

Elle peut aussi être réduite encore une fois en appliquant la règle qui construit ⟨rousse$_{(N)}$⟩. Cela nous donne la structure de la Figure 10a, puis celle en 10b une fois la règle de propagation des gouverneurs appliquée. Il résulte que les seuls sémantèmes accessibles à $\mathcal{I}_{\text{sém-synt}}$ (donc demandant à être exprimés), sont ⟨rousse$_{(N)}$⟩, ⟨habiter⟩ et ⟨Toronto⟩ ce qui permet de former l'énoncé (4).

(4) *Une rousse qui habite Toronto.*

On aurait pu réduire les sémantèmes autrement, en utilisant d'autres règles d'équivalence sémantique. Dans tous les cas, on retrouverait toujours la configuation sémantique de la Figure 6, mais de nouveaux sémantèmes réduits apparaîtraient et les sens accessibles à $\mathcal{I}_{\text{sém-synt}}$ne seraient pas les mêmes.

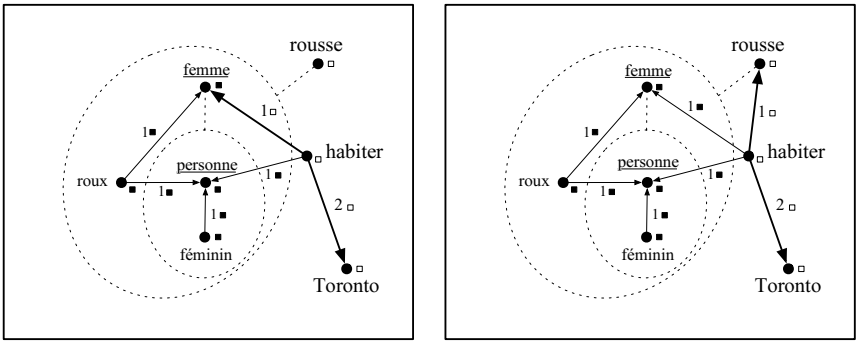


FIG. 10 – La réduction de ⟨femme rousse⟩ avec propagation des gouverneurs

5 Le typage des actants sémantiques

Le type de représentation sémantique que nous suggérons ici permet de représenter de façon assez naturelle le typage des actants sémantiques. Si on souhaite restreindre la combinatoire sémantique d'un prédicat, on peut en étiqueter les actants dans la règle de $\mathcal{G}_{\text{sém}}$ qui le construit. Par exemple, si on veut que le prédicat ⟨habiter⟩ ne puisse prendre comme premier actant qu'un sémantème dénotant une personne et comme deuxième actant un sémantème qui dénote un lieu, il suffit de modifier la règle (que nous avons déjà vue à la Figure 5) qui introduit ce sémantème de la façon illustrée à la Figure 11.

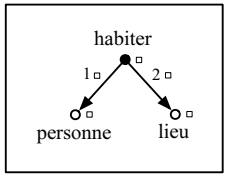


FIG. 11 – Le typage des actants sémantiques en GUST

Sans avoir accès à la décomposition du sens, le premier actant de ⟨habiter⟩ ne pourrait pas, par exemple, être ⟨rousse$_{(N)}$⟩ puisque l'étiquette ⟨personne⟩ du nœud en Figure 11 ne peut pas directement s'unifier avec l'étiquette ⟨rousse$_{(N)}$⟩. Par contre, avec les règles que nous avons présentées, puisque ⟨personne⟩ est le genre prochain de ⟨femme⟩, qui lui-même est le genre prochain de ⟨rousse$_{(N)}$⟩, ce dernier pourra, grâce à la règle de propagation des gouverneurs, être le premier actant de ⟨habiter⟩, sans qu'il ne soit nécessaire d'activer une quelconque procédure additionnelle ou de faire appel à une structure de donnée séparée qui encode une hiérarchie des types sémantiques. Comme on peut l'observer à la Figure 10b, le formalisme permet en effet de vérifier que ⟨personne⟩ est une composante centrale du premier actant de ⟨habiter⟩, et ce peu

importe à quelle « profondeur » se trouve enfouie cette composante dans le sens de l'actant. Bien entendu, puisque nous n'avons pas fourni dans nos règles la décomposition sémantique de ⟨Toronto⟩, le mécanisme ne pourrait pas fonctionner pour le deuxième actant et ⟨habiter⟩, tel que défini ici, ne pourrait pas se combiner à ⟨Toronto⟩. Mais si nous complétons notre grammaire pour fournir la décomposition de ce sémantème, ⟨habiter⟩ pourra l'accepter comme deuxième actant puisque ⟨Toronto⟩ ⊃ ⟨ville⟩ ⊃ ⟨lieu⟩.

Insistons sur le fait que le formalisme n'oblige pas à typer les actants d'un sémantème. On peut très bien ne pas étiqueter les actants dans les règles qui introduisent les sémantèmes prédicatifs sans que n'en souffre la grammaire. Le modèle permet alors de mettre en correspondance des phrases absurdes avec des représentations sémantiques formellement correctes, mais difficiles à interpréter. Revient-il à la grammaire de rejeter de tels énoncés ? D'un côté, il est vrai que la langue permet d'exprimer même des idées absurdes ou contradictoires. Si la grammaire n'est qu'un modèle fonctionnel de la langue, alors elle doit permettre de telles phrases. Le typage des actants risque, sans mécanisme approprié[6], de rendre la grammaire trop rigide pour permettre les écarts et glissements sémantiques. D'un autre côté, le typage des actants semble modéliser une certaine connaissance que les locuteurs ont des sens de la langue. De fait, dans le cadre de la TST, Mel'čuk et Polguère n'hésitent pas à y avoir recours dans leur base de données lexicales *DiCo* (Polguère, 2003). En outre, cela peut s'avérer utile pour diverses applications, notamment en linguistique computationnelle (par exemple pour valider une analyse incertaine). Enfin, la question dépasse le cadre de notre exposé, mais quoi qu'il en soit, le formalisme permet d'implémenter, au moins dans une certaine mesure, les deux solutions : le linguiste choisira à sa guise de typer ou non les actants des prédicats qu'il décrit dans la grammaire de bonne formation sémantique.

6 Conclusion

Le type de règle que nous proposons permet, sans aucune procédure particulière et sans avoir à ajouter ni nouveau module ni nouvelle polarité au modèle, de trouver automatiquement toutes les réductions possibles pour une représentation sémantique donnée, et donc toutes ses différentes expressions. Nous avons illustré le fonctionnement de ces règles en synthèse, mais elles fonctionnent également en analyse. Elles fournissent alors, pour l'énoncé analysé, non seulement sa représentation sémantique réduite, mais également sa décomposition complète ainsi que toutes les décompositions intermédiaires. Cette caractéristique du formalisme proposé peut s'avérer fort intéressante notamment en traduction automatique, puisque la traduction des lexèmes et des sens grammaticaux doit souvent être approximative, ce qui demande d'avoir accès à la décomposition des sens.

Remerciements

Nous tenons à remercier les relecteurs anonymes pour leurs commentaires constructifs ainsi que Dominique Longin pour son aide précieuse lors de l'édition de la version finale de cette article.

[6](Pustejovsky, 1995) propose justement de tels mécanismes.

Références

KAHANE S. (2001). Grammaires de dépendance formelles et théorie sens-texte : Tutoriel. In D. MAUREL, Ed., *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, volume 2, p. 17–76, Tours : ATALA.

KAHANE S. (2002). *Grammaire d'Unification Sens-Texte : vers un modèle mathématique articulé de la langue*. Thèse d'habilitation à diriger des recherches, Université Paris 7.

KAHANE S. (2004). Grammaires d'unification polarisées. In P. BLACHE, Ed., *Actes de TALN 2004 (Traitement automatique des langues naturelles)*, p. 233–242 : ATALA LPL.

KAHANE S. & LAREAU F. (2005). Grammaire d'unification sens-texte : modularité et polarisation. In *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, p. 23–32, Dourdan : ATALA LIMSI.

MEL'ČUK I. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary : basic principles and heuristic criteria. *International Journal of Lexicography*, **1**, 165–88.

MEL'ČUK I. (1989). Semantic primitives from the viewpoint of the meaning-text linguistic theory. *Quaderni di semantica*, **10**, 65–102.

MEL'ČUK I. (1997). *Vers une Linguistique Sens-Texte : Leçon inaugurale au Collège de France*. Paris : Collège de France.

MILIĆEVIĆ J. (sous presse). *Modélisation de la paraphrase langagière*. Bern : Peter Lang.

POLGUÈRE A. (2003). Étiquetage sémantique des lexies dans la base de données dico. *Traitement automatique des langues*, **44**(2), 39–68.

PUSTEJOVSKY J. (1995). *The generative lexicon*. Cambridge / London : The MIT Press.

WIERZBICKA A. (1996). *Semantics : primes and universals*. Oxford : Oxford University Press.

Systèmes de questions-réponses : vers la validation automatique des réponses

Anne-Laure LIGOZAT, Brigitte GRAU, Isabelle ROBBA, Anne VILNAT
LIMSI-CNRS - BP 133, 91403 Orsay Cedex
prenom.nom@limsi.fr

Résumé. Les systèmes de questions-réponses (*SQR*) ont pour but de trouver une information précise extraite d'une grande collection de documents comme le Web. Afin de pouvoir comparer les différentes stratégies possibles pour trouver une telle information, il est important d'évaluer ces systèmes. L'objectif d'une tâche de validation de réponses est d'estimer si une réponse donnée par un *SQR* est correcte ou non, en fonction du passage de texte donné comme justification. En 2006, nous avons participé à une tâche de validation de réponses, et dans cet article nous présentons la stratégie que nous avons utilisée. Celle-ci est fondée sur notre propre système de questions-réponses. Le principe est de comparer nos réponses avec les réponses à valider. Nous présentons les résultats obtenus et montrons les extensions possibles. À partir de quelques exemples, nous soulignons les difficultés que pose cette tâche.

Abstract. Question answering aims at retrieving precise information from a large collection of documents, typically the Web. Different techniques can be used to find relevant information, and to compare these techniques, it is important to evaluate question answering systems. The objective of an Answer Validation task is to estimate the correctness of an answer returned by a QA system for a question, according to the text snippet given to support it. We participated in such a task in 2006. In this article, we present our strategy for deciding if the snippets justify the answers. We used a strategy based on our own question answering system, and compared the answers it returned with the answer to judge. We discuss our results, and show the possible extensions of our strategy. Then we point out the difficulties of this task, by examining different examples.

Mots-clés : systèmes de questions-réponses, validation de réponses.

Keywords: question answering, answer validation.

1 Introduction

Les systèmes de questions-réponses (*SQR* par la suite) ont pour but de trouver une information précise dans une grande collection de documents. L'hypothèse sous-jacente au développement de tels systèmes est que les utilisateurs préfèrent en général recevoir une réponse précise à la question qu'ils se posent plutôt qu'un ensemble de documents à explorer, comme le proposent habituellement les moteurs de recherche (Voorhees, 1999). Cependant, pour être considéré comme fiable par un utilisateur, un *SQR* doit être capable de donner des éléments permettant d'évaluer ses réponses. L'objectif d'un système ne doit donc pas seulement être de trouver

les réponses, mais aussi de les exprimer d'une façon qui permette à l'utilisateur de savoir s'il peut avoir confiance en ces réponses. Ces éléments de justification donnent à l'utilisateur un moyen de vérifier que la réponse fournie correspond bien à l'information qu'il cherche, et ainsi de donner une valeur de vérité à cette réponse, en supposant que l'utilisateur a des connaissances « standard ».

Une bonne justification doit être concise et complète. Le but est de ne fournir que les extraits de documents qui permettent à l'utilisateur de retrouver toutes les informations qu'il a données, sans avoir à lire un document entier. Voici un exemple d'une telle justification.

> Question : Quand a eu lieu la chute du mur de Berlin ?
>
> Réponse : **en 1989**
>
> Justification (passage d'un document) : Cette ère de la dissuasion, fondée sur l'équilibre de la terreur entre deux grands blocs antagonistes, est remise en question **en 1989**, avec la chute symbolique du mur de Berlin.

2 Validation de réponses

(Lin & Pantel, 2001) soulignent la possible distance linguistique entre les questions et leurs réponses accompagnées de leur justification, en prenant l'exemple de la phrase *« Stendhal a écrit 'La chartreuse de Parme' en 1838 »* justifiant la réponse *« Stendhal »* à la question *« Qui est l'auteur de 'La chartreuse de Parme' ? »*. Ils définissent les liens entre une question et sa réponse justifiée par le terme d'*inférence*. Ils proposent alors de définir des *règles d'inférence* pour reconnaître par exemple la relation entre « X a écrit Y » et « X est l'auteur de Y ». Ces règles correspondent plus ou moins à ce qui est appelé *paraphrases* ou *variantes* dans d'autres travaux (Jones & Tait, 1984; Fabre & Jacquemin, 2000).

Le lien entre question et réponses correspond à la notion de *textual entailment* telle qu'elle est définie par Pascal Recognizing Textual Entailement Challenge [1] (RTE). *L'implication textuelle* est définie comme une tâche de décision qui à partir de deux fragments de texte, estime si d'un point de vue sémantique on peut déduire l'un de l'autre. Ainsi le passage de texte suivant (appelé *justification*) : *« Yoko Ono a inauguré une statue de bronze représentant son mari décédé, John Lennon, pour compléter le changement de nom officiel de l'aéroport de Liverpool qui devient l'aéroport John Lennon de Liverpool »* implique la phrase *« Yoko Ono est la veuve de John Lennon »* (appelée *hypothèse* dans le contexte de l'implication textuelle). Dans RTE, les participants reçoivent des paires justification-hypothèse de ce type et doivent ensuite décider si les hypothèses peuven ! t ou non être déduites des justification. Cette tâche est similaire à la tâche de réponses aux questions en ce qui concerne les questions booléennes (attendant *oui* ou *non* en réponse), car répondre à ces questions revient en fait à décider si la justification de la réponse implique la réponse.

En 2006, un nouvel exercice de validation des réponses, AVE [2], a été introduit dans la campagne de questions-réponses de CLEF. Le but de cet exercice est d'une part d'améliorer les performances des *SQR*, en développant des méthodes automatiques d'évaluation des réponses, et d'autre part de rendre le jugement humain semi-automatique à la condition que l'exercice produise des méthodes fiables d'évaluation. Pour cet exercice, les organisateurs ont produit un

[1] `http://www.pascal-network.org/Challenges/RTE`

[2] Answer Validation Exercise, `http://nlp.uned.es/QA/AVE/`

corpus à partir des réponses des participants à la tâche de questions-réponses et des passages de texte donnés comme justification. Les participants avaient alors pour tâche de décider pour chaque réponse si elle était correcte ou non en fonction du passage justificatif.

Les premiers travaux de validation automatique de réponses ont eu lieu au cours de la campagne AVE en 2006 ; cependant, les campagnes d'implication textuelle RTE avaient déjà proposé ce type de tâche.

Voici un exemple de couple (hypothèse, justification) d'AVE :

> Hypothèse : Yasser Arafat était **leader de l'Organisation de Libération de la Palestine** [3]
>
> Justification : Le président Clinton a fait appel personnellement au **leader de l'Organisation de Libération de la Palestine** Yasser Arafat et aux Palestiniens mercredi pour qu'ils reprennent les pourparlers en faveur de la paix avec Israël

Ici l'hypothèse est une reformulation de la question *« Qui était Yasser Arafat ? »* dans laquelle a été insérée une réponse proposée par un système *« leader de l'Organisation de Libération de la Palestine »*.

Dans AVE, le corpus de paires justification-hypothèse a été construit semi-automatiquement à partir des réponses obtenues par les participants lors de QA@CLEF 2006, campagne d'évaluation des *SQR*. Le corpus contient environ 3000 paires. Les participants à AVE ont été évalués sur leur capacité à prédire si une réponse (attestée par des juges humains) était correcte ou non. Ils avaient donc pour chaque paire deux possibilités de réponse : OUI ou NON.

Les résultats ont été évalués par la précision, le rappel et la f-mesure qui ont été calculés de la façon suivante :

$$précision = \frac{\#paires\ jugées\ OUI\ correctement}{\#jugées\ comme\ OUI}, rappel = \frac{\#paires\ jugées\ comme\ OUI\ correctement}{\#paires\ OUI}$$

$$\text{et } f\text{-}mesure = \frac{2*précision*rappel}{précision+rappel}$$

3 Travaux en validation de réponses

(Peñas *et al.*, 2006) présentent le déroulement de la première campagne AVE. 11 groupes ont participé à ce premier essai en soumettant 38 runs dans 7 langues différentes. L'anglais et l'espagnol étaient les langues les plus représentées avec respectivement 11 et 9 runs soumis. 2 groupes ont proposé des runs dans les 7 langues : ce sont les universités de Twente et d'Alicante.

Dans chaque langue, les paires *justification-hypothèse* ont été construites à partir des soumissions à la tâche questions-réponses de la campagne CLEF 2006. De ce fait, le pourcentage de paires positives, négatives et non évaluées [4] peut-être variable d'une langue à l'autre, ce qui ne permet pas réellement la comparaison des systèmes ayant participé dans des langues distinctes. Voici par exemple les pourcentages pour les 3 langues où les différences sont les plus importantes :

[3] Dans nos exemples, la réponse est écrite en **gras**.

[4] Les paires non évaluées de AVE proviennent de runs qui n'ont pu être évalués lors de la campagne QA@CLEF. En anglais et en portugais ce nombre est très élevé : 35% et 40%.

– en hollandais, OUI : 10%, NON : 86%, NON ÉVALUÉES : 4% ;
– en anglais, OUI : 10%, NON : 55%, NON ÉVALUÉES : 35% ;
– en espagnol, OUI : 28%, NON : 68%, NON ÉVALUÉES : 4% ;

Différentes approches ont été adoptées dans cette campagne. L'approche logique obtient les meilleurs résultats (Tatu *et al.*, 2006) [5], soulignons qu'elle est très souvent accompagnée de connaissances linguistiques : elles servent à transformer les éléments textuels en représentation logique. Au moins 3 équipes ont utilisé logique et connaissances linguistiques. Les approches qui utilisent de l'apprentissage sont également au nombre de 3 et l'une d'entre elle s'est attaquée aux 7 langues proposées. Elles utilisent des corpus déjà annotés comme ceux des campagnes RTE. Une approche, qui a participé elle aussi dans les 7 langues, adopte une méthode fondée sur les paraphrases : celles-ci sont engendrées automatiquement à partir de corpus bilingues alignés. Deux approches au moins utilisent des connaissances linguistiques sans faire référence à l'utilisation de la logique. Partant du constat qu'en espagnol 75% des questions de la campagne QA@CLEF étaient factuelles, une approche s'est fondée uniquement sur la reconnaissance d'entités nommés.

(Tatu *et al.*, 2006) utilisent un mécanisme de reconnaissance des entités nommées, un analyseur syntaxique et un analyseur sémantique pour transformer le passage justificatif et l'hypothèse en une représentation logique qu'ils qualifient de riche. Les représentations sont ensuite soumises à COGEX, qui détermine si oui ou non la justification implique l'hypothèse. La plupart des erreurs commises par ce système sont dues à une mauvaise syntaxe des hypothèses (celles-ci sont construites automatiquement), qui entraîne la construction de représentations logiques erronées. Néanmoins ce système obtient les meilleurs résultats dans les 2 langues dans lesquelles il a participé. En anglais, il obtient une f-mesure de 0.4393 et en espagnol une f-mesure de 0.6063.

(Ferrandez *et al.*, 2006) dérivent également une forme logique à partir du passage justificatif et de l'hypothèse. Pour cela, ils utilisent l'analyseur de Lin, MINIPAR (Lin, 2005), et obtiennent une représentation des phrases sous la forme d'un ensemble de relations de dépendances. Les relations sont ensuite transcrites dans des formes logiques, puis une mesure de similarité est calculée, celle-ci produit un poids sémantique utilisé pour juger si le passage justificatif implique ou non l'hypothèse. Ils ont soumis des runs dans toutes les langues et obtenu les meilleurs résultats en français (f-mesure : 0.4693) et en italien (f-mesure : 0.4066).

Pour leur participation à AVE, (Kouylekov *et al.*, 2006) ont adopté une approche fondée sur la notion de distance : ils essayent d'effectuer un *mapping* entre le contenu de l'hypothèse et la justification. Ils soulignent que plus ce *mapping* est direct plus il est probable que la justification implique l'hypothèse. Le *mapping* consiste ici en une séquence d'opérations d'édition, chacune ayant un coût. Les opérations (insertion, suppression, substitution) sont appliquées sur les arbres de dépendances du passage justificatif et de l'hypothèse. Quand le coût total de ces opérations est en dessous d'un seuil fixé, le passage justificatif est considéré comme impliquant l'hypothèse. Malgré différents problèmes dans la mise en place de ces modules, ils ont obtenu la 3ème place en anglais avec une f-mesure de 0.3776.

[5] Tous les articles évoqués dans ce paragraphe ne seront pas tous référencés, mais ils sont rassemblés dans les notes de travail du workshop CLEF 2006 et sont consultables à l'adresse `http://www.clef-campaign.org/2006/working_notes/CLEF2006WN-Contents.html`

Comme cela a été dit dans l'introduction, il existe une forte connexion entre AVE et RTE. La proposition à l'origine d'AVE était que l'on pouvait reformuler la tâche de validation de réponse comme un problème d'implication textuelle. Et, plusieurs groupes ont d'ailleurs participé aux deux évaluations en utilisant la même approche.

En 2006, a été organisé le second RTE. (Bar-Haim *et al.*, 2006) soulignent les particularités des deux systèmes qui ont obtenu les meilleurs résultats. L'un a utilisé de façon extensive des connaissances sémantiques, l'autre a favorisé l'utilisation de grands corpus d'entraînement.

Dans notre travail, nous ne faisons pas l'hypothèse d'une source de connaissances sémantiques existante qui permettrait des déductions logiques. Aussi, nous reposons nous sur des critères linguistiques, qui peuvent être vérifiés en domaine ouvert, et qui permettent d'exprimer des relations sémantiques entre le sens des mots.

4 Valider des réponses avec un *SQR*

Notre objectif était d'utiliser notre propre *SQR* pour le français : FRASQUES, et d'utiliser ses résultats, c'est-à-dire à la fois les réponses extraites et les types d'informations de la questions présentes dans les justifications, pour évaluer la pertinence des justifications par rapport aux hypothèses.

4.1 FRASQUES : notre système de questions-réponses pour le français

Nous présentons tout d'abord brièvement FRASQUES avant de présenter comment il a été adapté pour la tâche de validation.

Le système se divise en 4 composants :
- Analyse de la question : ce premier module effectue l'analyse syntaxique de la question pour en détecter certaines de ses caractéristiques telles que :
  - ses mot-clés, utilisés ultérieurement lors de la recherche des documents,
  - le type attendu de la réponse, qui peut-être une entité nommée (une personne, un pays, une date...) ou un type général comme *conférence* ou *adresse*,
  - le focus de la question, que nous définissons comme le terme de la question qui sera vraisemblablement présent dans la phrase contenant la réponse,
  - le verbe principal de la question.
- Sélection des documents : le moteur de recherche Lucene [6] cherche dans la collection les documents pertinents.
- Traitement des documents : ce module utilise Fastr [7] pour reconnaître les variantes linguistiques des termes de la question : par exemple, « monnaie de l'Europe » sera reconnue comme une variante de « monnaie européenne ». Ensuite, les entités nommées du document sont étiquetées, nous utilisons environ une vingtaine de type d'entités nommées. Les phrases contenant au moins une variante des termes de la question sont gardées.
- Extraction de la réponse : ce dernier module extrait les réponses précises des phrases candidates. La stratégie d'extraction dépend du type attendu de la réponse. Si la réponse est une entité nommée, l'entité nommée qui est du type attendu et qui est la plus proche des mots de

[6] Moteur de recherche entièrement écrit en Java `http://lucene.apache.org/`
[7] `http://www.limsi.fr/Individu/jacquemi/FASTR/`

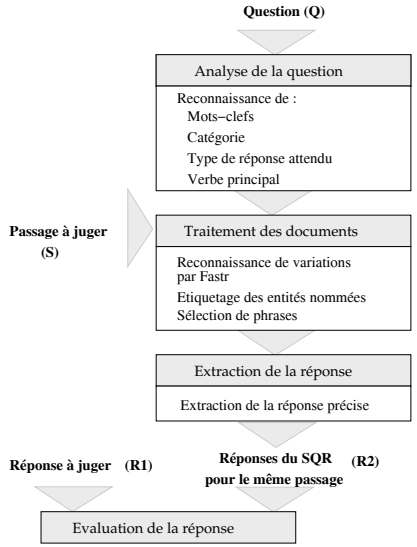


FIG. 1 – Architecture du système de validation de la réponse

la question est sélectionnée. Sinon, des patrons d'extraction sont utilisés, ils sont écrits dans le format Cass [8], un analyseur syntaxique qui est utilisé ici pour extraire la réponse plutôt que comme analyseur. Ces patrons expriment la position possible de la réponse par rapport au focus ou au type attendu de la réponse.

4.2 Le système de validation des réponses

Le système de validation des réponses utilise trois de ces quatre composants, ce que montre la figure 1. L'entrée du système est une paire justification-hypothèse, ainsi que la question d'origine *Q* et la réponse à juger *R1*. La question est d'abord analysée puis le composant qui traite les documents est appliqué à la justification. Le module d'extraction de la réponse extrait les réponses *R2* des passages justificatifs. Enfin, la paire hypothèse-justification est évaluée en tenant compte des différentes informations de l'hypothèse trouvées dans l'extrait et de la réponse trouvée par FRASQUES. Le système retourne OUI si elle est considérée comme justifiée, NON dans le cas contraire. Un score de confiance est également attribué à chaque jugement.

L'algorithme de décision se déroule en 2 étapes. La première a pour but de détecter les erreurs les plus triviales, par exemple une réponse qui serait complètement incluse dans la question ou qui ne serait pas présente dans la justification. Dans le cas où la question contient une date, le contexte temporel de la question et l'extrait sont comparés. Pour l'instant, le contexte est formé par les dates reconnues comme telles présentes dans la description du document ou dans le passage. S'ils sont contradictoires, la paire est rejetée.

[8] `http://www.sfs.nphil.uni-tuebingen.de/~abney/`

La seconde étape consiste en des vérifications plus complexes. Dans un cas idéal, un passage justificatif correct correspond à la reformulation de la question sous forme déclarative avec la réponse qui y est donnée. Chaque terme de la question, ou de l'hypothèse, figure dans le passage, liés par les mêmes relations.

En ce qui concerne les termes, dans la grande majorité des cas, le passage justificatif ne comporte pas tous les termes de la questions sous leur forme d'origine : ils subissent des variations de différentes natures : flexionnelles, morphologiques, syntaxiques, sémantiques ou des combinaisons de ces variations si on recherche des groupes nominaux complexes. Dans FRASQUES, ces variations sont reconnues par Fastr. Parmi les termes de la question, certains jouent un rôle plus important. Il en est ainsi de l'objet de la question, que nous appelons focus dans FRASQUES. Le focus correspond à l'entité sur laquelle porte la question, que l'on en cherche une caractéristique ou une définition. Aussi, selon les types de question, le focus n'est pas toujours présent, mais s'il l'est, il doit figurer dans le passage justificatif. Un autre terme qui, s'il est présent, a une grande importance, est le type de réponse attendu, quand ce type n'est pas un nom d'entité nommée. Ce type est nommé type général. Ainsi, dans « De quel parti politique Lionel Jospin est-il membre ? » Le focus est « Lionel Jospin » et le type général est « parti politique ». Lorsqu'il est présent dans le passage réponse, le type général sera souvent placé à proximité de la réponse ou même fera partie de celle-ci, comme dans « Lionel Jospin, membre du parti socialiste ».

Lorsqu'il s'agit du verbe, celui-ci a tendance à subir plus de variations que les termes nominaux ; il est souvent exprimé par une préposition ou un verbe proche mais non synonyme. C'est le cas par exemple si on demande « qui a réalisé un film » et que la réponse est exprimée par « le film de X ... » ou « quelle entreprise a changé son nom » et la réponse est donnée par « le groupe X a adopté le nom de la filiale ... ». On retrouve ici les variations traitées par (Lin & Pantel, 2001). Ne disposant pas de telles ressources, nous avons considéré que l'absence du verbe n'influerait pas sur la décision finale.

Enfin les derniers types de termes jouant un rôle primordial sont les noms propres : ils sont toujours présents dans le passage et subissent peu de variations, sauf en ce qui concerne les noms de pays souvent repris par l'adjectif correspondant, comme dans « qui est le président de l'Egypte » avec « le président égyptien » repris dans le passage.

En ce qui concerne les relations entre termes, celles-ci seront souvent vérifiées par leur manifestation en langue, c'est-à-dire par un ensemble de relations syntaxiques. Nous avons vu que certains travaux s'appuient sur une notion de distance syntaxique. Mais pour cela, il est nécessaire de disposer d'une analyse complète des phrases. Afin de ne pas reposer sur cette hypothèse souvent non vérifiée, nous avons choisi de ne vérifier que certaines relations en les exprimant sous forme de patrons d'extraction. Ces relations sont celles qui lient la réponse avec certains éléments de la phrase : le focus ou le type général.

L'élément prépondérant, malgré tout, reste la réponse : est-elle du type attendu ou non ? Lorsque ce type est une entité nommée, la vérification consistera à retrouver une entité nommée d'un type adéquat. Lorsque celui-ci est désigné par le type général, ou bien il figure à proximité de la réponse, ou bien il est implicite et la réponse en est une instance. Cette relation d'instanciation pourrait être inférée par l'utilisation de ressources externes, par exemple Wikipedia, qui possède un grand nombre de catégories et de définitions leur correspondant.

La mise en oeuvre de ces critères de justification donne lieu dans notre système à un calcul de 2 scores qui permet ensuite de conclure positivement ou négativement. Le premier score

porte sur l'évaluation de la correspondance entre la réponse trouvée par notre système *R2* et la réponse proposée dans l'hypothèse *R1*. Si FRASQUES trouve une réponse différente, alors la paire hypothèse-justification est réfutée. Si les 2 réponses sont proches ou s'il n'y a pas de réponse trouvée par FRASQUES, la décision va être conditionnée par la présence des différents termes que nous avons privilégiés. Le score attribué à l'évaluation de la qualité de la réponse sera positif pour une réponse exacte ou approchée, et négatif quand la distance est assez grande, par exemple, l'approximation d'une date ou d'une quantité par un nombre.

Le deuxième score évalue les termes présents. Il est calculé en combinant le nombre de critères présents et leurs valeurs. Il est négatif si aucun des critères n'est trouvé, et positif sinon.

Un passage constitue une justification acceptable :
– si *R2* est absente et le score des termes est positif. Ce dernier fournit le score final,
– si *R2* = *R1* et il y a des critères présents. Dans ce cas le score final est la valeur maximale des deux critères,
– si les 2 scores vont dans des sens opposés, on prend le meilleur des deux, s'il est positif.
Regardons l'exemple suivant :

> Justification : Trois candidats, Tony Blair, Margaret Beckett et John Prescott, se disputeront la succession de John Smith à la tête du parti travailliste, a annoncé le **Labour** jeudi, à l'issue du processus de nominations des candidats par les députés du parti. M. Blair, ministre de l'Intérieur du cabinet fantôme représentant l'a
>
> Hypothèse : le parti politique de Tony Blair, le **LABOUR** .

Dans ce passage, tous les termes de la question sont présents (*Tony Blair, parti politique*), mais le type de la réponse *Labour* n'est pas étiqueté par notre *SQR* comme une entité nommée de type organisation. De ce fait, l'algorithme de décision reçoit 2 scores opposés, dans cet exemple prenant en compte le non-marquage de *Labour* comme une organisation, il répond négativement.

4.3 Résultats

Le corpus d'évaluation contenait 3266 paires, parmi lesquelles 202 paires n'ont pas été jugées. Les hypothèses étant formées automatiquement, elles comportaient beaucoup d'erreurs de syntaxe, aussi nous sommes-nous fondés uniquement sur les questions, l'hypothèse ne nous permettant que d'extraire la réponse.

Lors de notre participation à AVE, beaucoup d'erreurs restaient dans nos programmes, qui ont été corrigés depuis. Les organisateurs ayant fourni les valeurs de validation attendues pour chaque paire du corpus, nous avons pu réévaluer notre chaîne. Un examen approfondi de ces résultats nous a permis de constater qu'il y avait certaines erreurs sur ces valeurs, notamment en ce qui concerne les réponses positives : des réponses exactes aux questions n'étaient pas du tout validées par le passage justificatif. Nous avons corrigé 82 d'entre elles dans le corpus.

La table 1 présente nos résultats sur la version officielle, les corrections apportées au corpus ne modifiant pas les ordres de grandeur des résultats. La première ligne donne le nombre de paires évaluées positivement et négativement par les juges humains. La seconde ligne contient tous nos résultats et la suivante le nombre de nos résultats corrects. La dernière ligne contient le rappel et la f-mesure correspondant à ces résultats en utilisant la formule exposée dans la section 2.

	# OUI	# NON	Total
Évalués par les organisateurs	705	2359	3064
Tous nos résultats	142	2922	3064
Nos résultats corrects	82	2266	2348
Précision	0.58	0.77	
Rappel	0.12	0.96	
F-mesure	0.2	0.85	

TAB. 1 – AVE results at CLEF 2006

Parmi nos réponses NON, nous distinguons celles qui sont sûres des autres : ce sont les réfutations décidées lors de la première étape présentées dans la section 4.2. La réponse est considérée comme étant non justifiée et ce de façon sûre, donc avec un score de confiance élevé. Nous avons trouvé 1637 paires de « NON » sûrs. Parmi elles, 1415 étaient bien jugées, la précision pour ces réponses est donc de 0,87.

La seconde observation est que notre système a plus de facilités pour réfuter les justifications plutôt que pour les accepter. La précision et le rappel de nos réponses négatives sont bons. Et nous nous trompons rarement quand nous donnons des réponses OUI, mais nous en trouvons très peu, notre rappel est donc très faible sur ces réponses.

Certaines erreurs pourraient être corrigées en approfondissant les vérifications des relations portant sur la réponse. Par exemple, pour la question *« Quel est le nom de la femme de George W. Bush ? »*, une des hypothèses construites était *« **Norman Schwarzkopf**, la femme de George W. Bush. »*. On pourrait alors interroger le Web avec la requête *femme de George W. Bush* et constater que la ou les réponses obtenues sont fortement incompatibles avec ***Norman Schwarzkopf***.

5 Conclusion

Nous avons présenté une stratégie de validation des réponses issues d'un *SQR*. Cette stratégie est fondée sur FRASQUES, notre propre *SQR* monolingue : l'hypothèse et l'extrait sont analysés par FRASQUES et nous utilisons des critères qui permettent de détecter si l'extrait justifie ou non la réponse. Dans notre évaluation des paires hypothèses-extrait, nous distinguons avec une bonne précision les cas dans lesquels l'extrait ne justifie pas la réponse. Des possibilités d'extension de notre stratégie, utilisant des ressources externes et nous permettant d'acquérir de nouvelles connaissances ont également été présentées.

Cette première expérience en validation de réponses constitue une étape vers la validation semi-automatique en questions-réponses. Elle nous permettra à terme d'améliorer les performances de notre *SQR* puisque certains des critères que nous utilisons pour la validation n'y avaient pas été mis en œuvre.

Références

BAR-HAIM R., DAGAN I., DOLAN B., FERRO L., GIAMPICCOLO D., MAGNINI B. & SZPEKTOR I. (2006). The second pascal recognising textual entailment challenge. In *The Second*

PASCAL Challenges Workshop on Recognising Textual Entailment.

FABRE C. & JACQUEMIN C. (2000). Boosting Variant Recognition with Light Semantics. In *Proceedings of 18th International Conference on Computational Linguistics (COLING-2000)*, Sarrebrück, Allemagne.

FERRANDEZ O., TEROL R. M., MUNOZ R., MARTINEZ-BARCO P. & PALOMAR M. (2006). A knowledge-based textual entailment approach applied to the qa answer validation at clef 2006. In *Workshop CLEF 2006*, Alicante, Spain.

JONES K. S. & TAIT J. I. (1984). Automatic Search Term Variant Generation. *Journal of Documentation*, p. 50–66.

KOUYLEKOV M., NEGRI M., MAGNINI B. & COPPOLA B. (2006). Towards entailment-based question answering : Itc-irst at clef 2006. In *Workshop CLEF 2006*, Alicante, Spain.

LIN D. (2005). Dependancy-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems*, Southampton, UK.

LIN D. & PANTEL P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering*, **7**(04), 343–360.

PEÑAS A., RODRIGO A., SAMA V. & VERDEJO F. (2006). Overview of the answer validation exercise 2006. In *Workshop CLEF 2006*, Alicante, Spain.

TATU M., ILES B. & MOLDOVAN D. (2006). Automatic answer validation using cogex. In *Workshop CLEF 2006*, Alicante, Spain.

VOORHEES E. M. (1999). TREC-8 Question Answering Track Evaluation. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)* : Department of Commerce, National Institute of Standards and Technology.

Ressources lexicales chinoises pour le TALN

Huei-Chi LIN, Max SILBERZTEIN
Laboratoire de Sémiolinguistique, Didactique, Informatique (LASELDI) –
Université de Franche-Comté, 30 rue Mégevand 25000 Besançon
lin_huei_chi@yahoo.fr
max.silberztein@univ-fcomte.fr

Résumé. Nous voulons traiter des textes chinois automatiquement ; pour ce faire, nous formalisons le vocabulaire chinois, en utilisant principalement des dictionnaires et des grammaires morphologiques et syntaxiques formalisés avec le logiciel NooJ. Nous présentons ici les critères linguistiques qui nous ont permis de construire dictionnaires et grammaires, sachant que l'application envisagée (linguistique de corpus) nous impose certaines contraintes dans la formalisation des unités de la langue, en particulier des composés.

Abstract. In order to parse Chinese texts automatically, we need to formalize the Chinese vocabulary by using electronic dictionaries and morphological and syntactic grammars. We have used the NooJ software to enter the formalization. We present here the set of linguistic criteria used to construct these dictionaries and grammars, so that they can be used by corpus-linguistic applications. We focus our discussion on the characterization of Chinese linguistic units, specifically compounds.

Mots-clés : ressources linguistiques pour le chinois, linguistique de corpus, NooJ.
Keywords: linguistic resources for chinese, corpus linguistics, NooJ.

1 Introduction

Notre but est de formaliser le vocabulaire de la langue chinoise, plus précisément le mandarin tel qu'on le trouve dans les textes littéraires à partir du XX^{e} siècle, et écrit avec les caractères chinois traditionnels codés avec UNICODE (ce qui représente 70.207 caractères) pour les besoins des applications de linguistique de corpus et d'analyse syntaxique automatique. Nous avons donc entrepris la construction d'un module chinois pour NooJ[1]. Ce travail de recherche nous a conduits à construire des dictionnaires électroniques, des grammaires morphologiques et des grammaires syntaxiques. Toutes les unités du vocabulaire chinois doivent être recensées

[1] Cf. http://www.nooj4nlp.net. NooJ ainsi que le module chinois et d'autres ressources peuvent être téléchargés librement, et les utilisateurs de NooJ peuvent développer leurs propres ensembles de ressources linguistiques pour formaliser divers niveaux des langues : orthographe, lexique, morphologie, syntaxe et sémantique. Ces ensembles de ressources peuvent être rassemblées dans un « module » autonome, qui peut ensuite être chargé par d'autres utilisateurs pour analyser des textes de grande taille.

systématiquement et décrites explicitement. Il est impératif de formaliser tous les types d'unités linguistiques, et pas simplement les mots simples, et aussi de décrire leurs variations lexicales et morphologiques, ce que n'ont pas fait systématiquement les dictionnaires traditionnels chinois jusqu'à présent. Par exemple, le dictionnaire 辭海 (Cíhǎi) contient l'entrée lexicale « 畫冊 » (huàcè) [album de peintures], mais ne la relie pas à sa variante orthographique « 畫冊 » (huàcè) [album de peintures], ni à sa variante morphologique « 畫冊兒 » (huàcèér) [album de peintures]. En chinois, il n'y a pas de blanc séparateur de mots dans les textes. La reconnaissance automatique des mots chinois doit donc passer par la consultation de dictionnaires électroniques ou de grammaires morphologiques ou syntaxiques complets ; ce problème ressemble beaucoup à celui de la reconnaissance des mots composés et expressions figées dans les langues romanes, où l'on ne sait pas a priori où s'arrête un mot composé, et où seule un recensement systématique et une description précise permettent de distinguer les mots composés lexicalisés (par ex. « carte bleue ») des séquences libres de mots (« carte marron »). Nous avons dû adopter une série de critères linguistiques précis et surtout reproductibles, pour décider si une séquence de caractères chinois doit ou non être lexicalisée. Ces critères sont différents de ceux utilisés en linguistique traditionnelle : par exemple, dans le dictionnaire 辭海 (Cíhǎi), on trouve des entrées telles que « 白吃 » (báichī) [prendre gratuitement un repas], que nous n'avons pas de raison de lexicaliser, tandis que nous décrivons explicitement l'entrée « 鋼琴家 » (gāngqínjiā) [pianiste], non répertoriée dans le dictionnaire 辭海 (Cíhǎi).

Nous avons développé le module chinois pour NooJ afin d'analyser automatiquement des textes, de façon similaire aux systèmes CKIP et ICTCLAS ; notre module permet également d'identifier les entités nommées, et d'extraire à volonté des motifs syntactico-sémantiques à partir de requêtes d'utilisateurs. Notre module n'a pas d'application directe en traduction automatique, ni pour la comparaison entre le chinois, le japonais et le coréen. Enfin, nos dictionnaires électroniques ne contiennent pas de synonymes (au contraire de WordNet).

2 Définition des unités

Les Unités Linguistiques Atomiques (*Atomic Linguistic Units*, ou ALUs) de NooJ constituent les unités les plus petites de la langue qui doivent être associées à des informations linguistiques. Formellement, NooJ traite les ALUs en quatre classes formelles distinctes :

1) **Affixes** (préfixe, infixe et suffixe) : Ce sont des séquences de caractères chinois décrites par un composant morphologique, ou qui interviennent dans des opérations lexicales de flexion ou de dérivation. Par exemple, le préfixe « 初 » (chū) et le suffixe « 兒 » (ér)[2].

2) **Mots simples** : NooJ traite a priori chaque caractère chinois comme un mot simple. Par exemple, le caractère « 樹 » (shù) [arbre] constitue un mot simple et est utilisé dans les textes dans des contextes libres.

3) **Mots composés** : Ce sont des séquences de caractères que nous devons lexicaliser, comme par exemple « 蝴蝶 » (húdié) [papillon] et « 嘰哩咕嚕 » (jīlīgūlū) [avoir faim].

[2] 初 (chū) est un préfixe déterminatif que l'on place devant les dix premiers numéros de jours des mois, ce qui correspond au « **er** » dans « le 1er janvier ». 兒 (ér) est un suffixe phonétique pur (qui ne change pas le sens des mots), qu'on utilise après certains noms, verbes, adjectifs ou adverbes.

4) **Expressions figées** : Ce sont des séquences de caractères potentiellement discontinues. Par exemple, à l'intérieur de l'expression « 拖下水 » (tuō xiàshuǐ) [implanter dans l'eau — impliquer quelqu'un dans une situation], on peut insérer des pronoms personnels comme « 我 » (wǒ) [je], « 你 » (nǐ) [tu], « 他 » (tā) [il], etc. :

你要**拖**他**下水**。 (nǐ yào **tuō** tā **xiàshuǐ**) [Tu veux l'impliquer dans cette situation.]

NooJ traite différemment les affixes, mots simples et mots composés, qui sont des séquences insécables constitués d'un ou de plusieurs caractères, des expressions figées qui elles peuvent être discontinues. Il s'agit donc d'une part d'intégrer et de décrire les quatre types d'ALUs du chinois dans des dictionnaires : d'affixes, de mots simples, de mots composés et d'expressions figées ; et d'autre part, de construire des grammaires qui permettent de décrire les conditions d'utilisation et de combinaison de ces ALUs.

Par rapport aux approches lexicographiques traditionnelles, les expressions idiomatiques, les mots polymères chinois (cf. ci-dessous) et les mots surcomposés « en paire » (mots de 2x2 caractères) ont été simplement intégrés à nos dictionnaires puisque ce sont des ALUs comme les autres du point de vue du TALN. Par ailleurs, certaines ALUs ont été formalisées selon deux méthodes : l'une est de les ranger directement dans un dictionnaire ; l'autre est de les représenter à l'intérieur de grammaires locales, morphologiques ou syntaxiques.

Dans l'alphabet chinois traditionnel (au contraire de l'alphabet chinois simplifié), il existe de nombreuses variantes orthographiques, i.e. lorsqu'un mot ou un morphème s'écrit avec deux orthographes ou plus, et se prononce de la même façon dans tous les cas. Plusieurs variantes d'un même caractère peuvent cohabiter dans un même texte. Nous avons donc entré une table d'équivalence qui contient plus de 1.000 paires, telles que 巃 = 龍 et 悪 = 惡, et nous avons modifié l'algorithme de consultation des dictionnaires de NooJ pour qu'il prenne en compte ces équivalences.

3 Critères de lexicalisation

Nous présentons une série de critères qui nous ont permis de formaliser le vocabulaire chinois. Pour ce faire, nous avons dû adapter les critères utilisés avec NooJ pour décrire les ALUs des langues romanes (présentées dans Silberztein 1993) ; de plus il a fallu résoudre les difficultés spécifiques au chinois, principalement :

1) Du point de vue orthographique, la reconnaissance automatique des ALUs chinoises est plus complexe que celle des ALUs pour les langues romanes : un caractère chinois peut soit correspondre à une ALU autonome (comme un « mot simple » français), soit à un composant d'un ensemble d'ALUs productifs (« préfixe » que l'on traite avec des règles morphologiques), soit un composant d'une ALU plus longue (comme un « mot composé ») (Lin 2006).

2) En chinois, bien plus que dans les langues romanes et l'anglais, les mots ont systématiquement plusieurs fonctions syntaxiques. Par exemple, le même mot « 解釋 » (jiěshì) [expliquer ou explication] peut être indifféremment un verbe ou un nom :

他向我**解釋**他昨天缺席的原因。 [Il m'**explique** pourquoi il a été absent hier.]

我接受了他的**解釋**。 [J'ai accepté son **explication**.]

Les ambiguïtés concernent aussi les adjectifs et les adverbes :

他**和藹可親**。 [Il est gentil.] 他**和藹可親**地與我說話。 [Il m'a parlé gentiment.]

les adjectifs et les noms, etc. Pour calculer la fonction d'un mot dans la phrase (ce qui correspond à l'étiquetage pour les langues romanes), il faut donc bien souvent analyser syntaxiquement préalablement la phrase complète. Il est donc impossible de commencer l'analyse d'un texte par une étape d'étiquetage et de levée d'ambiguïtés : l'analyseur lexical des textes chinois produira donc nécessairement un résultat massivement ambigu, qui sera transmis à l'analyseur syntaxique. L'architecture spécifique de NooJ, qui permet d'étiqueter les textes partiellement ambigus en produisant une structure d'annotations potentiellement ambiguë, est donc bien adaptée à l'analyse lexicale des textes chinois.

3.1 Compositionnalité

La majorité des ALUs chinoises sont constituées d'au moins deux caractères. Parmi ces mots, certains ne peuvent pas être décomposés car ils contiennent des caractères non-autonomes, i.e. qu'on ne trouve nulle part ailleurs que dans ces mots. Considérons les mots ci-dessous :

蝴蝶 (húdié) [papillon], 嘰哩咕嚕 (jīlīgūlū) [avoir faim]

Ces deux mots contiennent des caractères qui n'ont pas d'utilisation autonome, par exemple, on ne trouve pas le caractère « 蝴 » en dehors du mot [papillon], et on ne trouve pas le caractère « 嘰 » en dehors de [avoir faim]. Il faut donc recenser et décrire ces mots dans un dictionnaire.

En revanche, certains mots peuvent être constitués de caractères qui peuvent avoir d'autres emplois de façon indépendante. Par exemple, considérons le mot suivant :

白菜 (báicài) [chou chinois]

Ce mot est constitué du caractère « 菜 » (qui signifie « légume ») et du caractère « 白 » (qui signifie « blanc »). Mais le sens du mot [chou chinois] ne peut pas être déterminé à partir du sens de ses deux constituants. En conséquence, il faut absolument lexicaliser ce mot, d'une part pour obtenir la bonne analyse du mot, et d'autre part, pour bloquer une analyse compositionnelle qui produirait le résultat incorrect « légume blanc » (un légume blanc s'écrirait « 白的菜 »).

3.2 Institutionnalisation

Beaucoup de concepts ou objets du monde réel sont désignés ou nommés systématiquement de la même façon par les locuteurs d'une langue, et ce de façon arbitraire. Par exemple, le concept [cœur sensible] s'exprime en chinois par le terme « 豆腐心 », littéralement « cœur de tofu », alors que d'autres expressions très semblables, telles que « 牛奶心 » (« cœur de lait ») ou « 白紙心 » (« cœur de purée ») ne peuvent pas utilisées pour exprimer ce concept.

Les formes dont l'usage est « institutionnalisé » ne sont pas morphologiquement, syntaxiquement ou sémantiquement différentes des autres formes potentielles qui ne sont jamais utilisées par les locuteurs chinois. Il faut donc distinguer formellement les termes vraiment employés par les locuteurs chinois des expressions potentielles qui ne le sont jamais.

Dans certaines applications de la formalisation des langues, telles que la traduction automatique ou l'enseignement des langues secondes, il est nécessaire de lexicaliser ces termes, qui correspondent à la « bonne » traduction, ce qui permet d'éviter des fautes d'analyse ou de traduction. Ainsi, pour traduire correctement l'expression française « cœur sensible », il faut produire « 豆腐心 » sans chercher à analyser les deux constituants de l'expression française ou de sa traduction : on traduit donc le tout « en bloc », ce qui revient à dire que l'expression complète est une ALU lexicalisée. La comparaison systématique entre les termes et des expressions similaires potentielles montre d'une part, que leur structure syntaxique n'est pas spécifique, d'autre part que la construction des termes ne peut pas être calculée par des règles morphologiques, syntaxiques ou sémantiques.

3.3 Structure des mots composés

A chaque fois que certaines propriétés syntaxiques ou sémantiques d'une forme chinoise ne peuvent pas être calculées à partir de celles de ses constituants, il faut lexicaliser cette forme et la traiter en tant qu'ALU, i.e. « en bloc ». Cependant, on ne peut pas ne pas noter que de nombreux termes composés chinois se construisent selon quelques schémas productifs. Nous décrivons ces schémas.

3.3.1 Mots polymères

Les mots polymères se composent d'au moins quatre caractères potentiellement autonomes dont les sens sont « similaires ». Sémantiquement, les polymères chinois sont construits par des mécanismes relevant de la coordination. Les caractères constituant un mot polymère ne peuvent pas être remplacés par d'autres caractères. Cependant, leur ordre d'apparition peut être modifié à l'intérieur du polymère. Par exemple, le terme « 紙墨筆硯 » qui signifie [trousse] peut s'écrire de huit façons différentes :

紙墨筆硯 (zhǐ mò bǐ yàn) [papier, encre, pinceau, encrier], mais aussi :

紙筆硯墨 紙筆墨硯 紙硯筆墨 筆硯紙墨 筆墨硯紙 筆墨紙硯 硯墨紙筆

80% des mots polymères se composent de quatre caractères. Les polymères sont fréquemment employés dans des textes. On distingue traditionnellement trois sortes de polymères :

1) Les quatre caractères ont des sens semblables. Par exemple :

酸甜苦辣 (suāntián kǔlà) [aigre, doux, amer, âcre → aléas de la vie]

Les quatre caractères appartiennent au même champ sémantique.

2) Les quatre caractères sont semblables deux à deux ; les deux premiers forment une paire, et les deux derniers forment une autre paire. Par exemple :

兄弟姊妹 (xiōngdì jiěmèi) [grand frère, petit frère, grande sœur, petite sœur → frères et sœurs]

姊妹兄弟(jiěmèi xiōngdì) [grande sœur, petite sœur, grand frère, petit frère → frères et sœurs]

Ce terme désigne l'ensemble des enfants d'une même famille.

3) Les deux premiers caractères qualifient les deux derniers. Par exemple :

金銀珠寶 (jīnyín zhūbǎo) [or, argent, objet d'une grande valeur → trésor]

Ici il n'y a pas de possibilité de permutation des caractères. « 金 » (jīn) et « 銀 » (yín) appartiennent à la même classe sémantique : métaux précieux, et qualifient le mot « 珠寶 » (zhūbǎo), qui représente la deuxième partie de ce polymère.

3.3.2 Mot radical + suffixe signifiant

Certains termes sont constitués d'un suffixe signifiant associé à un ensemble spécifique de mots de distribution restreinte. Par exemple, le mot « 團 » (tuán) [groupe] peut être combiné avec une centaine de mots selon la règle morphologique productive chinoise : **mot radical + Suffixe signifiant** :

合唱團 (héchàngtuán) [groupe de chanteurs = chorale]

訪問團 (fǎnwèntuán) [groupe d'interviewers = équipe journalistique]

觀光團 (guānguāngtuán) [groupe de touristes]

tandis que d'autres suffixes (匠 (jiàng), 員 (yuán), 商 (shāng), 家 (jiā), etc.) s'utilisent avec d'autres mots. Plutôt que de lexicaliser toutes ces formes dans un dictionnaire, il vaut mieux construire une grammaire locale NooJ qui les décrit de façon unifiée, cf. Figure 1 :

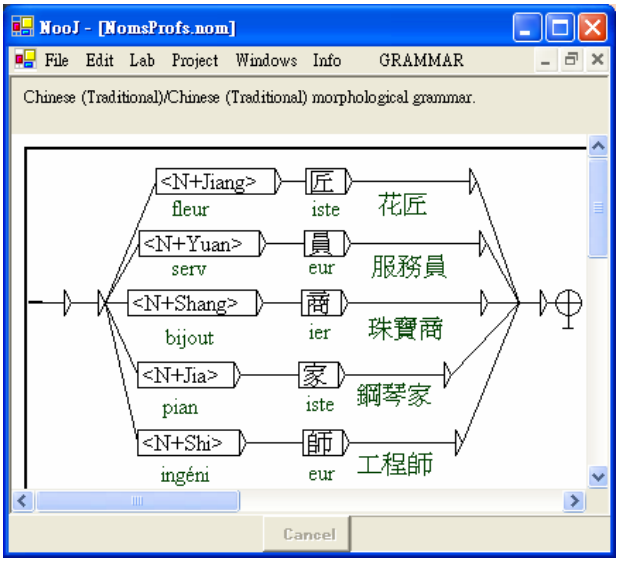


Figure 1 : Une grammaire locale de termes composés

Dans les grammaires de NooJ, les symboles comme <N+Jiang> réfèrent à des informations lexicales et représentent tous les noms (N) associés à la propriété « **+Jiang** » (qui ont été marqués comme pouvant être suivis du suffixe 匠). Noter que la description en extension de

ces ALUs dans une grammaire locale représente le fait que leur mode de construction est productif, mais pas que ces termes sont analysables : cette situation se retrouve en français, où par exemple les termes en « assurance » sont productifs : *assurance maladie*, *assurance vie*, *assurance chômage*, etc. mais chacun des termes construits est non-analysable (par ex. une *assurance vie* n'est pas une assurance contre la vie).

3.3.3 Mots composés de structure **XY**

En chinois, certains mots composés peuvent être créés sans conjonction. Ces mots composés sont classés en trois catégories et doivent être aussi rangés dans un dictionnaire :

1) Les deux constituants sont des lemmes lexicaux semblables. Par exemple :

金漿玉醴 (jīnjiāng yùlǐ) [liqueur d'or + vin de jade → un vin délicieux]

Il s'agit donc d'un mot composé de deux mots « 金漿 » et « 玉醴 » de même classe sémantique qui se combinent.

2) Les deux constituants sont synonymes. Par exemple :

公子哥兒 (gōngzǐ gēér) [jeune homme, garçon → fils de riche]

Ici, le deuxième constituant est un synonyme du premier. En chinois, beaucoup de termes sont construits grâce à des répétitions synonymiques. Ici, les deux constituants « 公子 » [jeune homme] et « 哥兒 » [garçon] sont obligatoirement juxtaposés, et ne peuvent pas être utilisés pour construire le sens de « fils de riche » dans une phrase s'ils sont isolés l'un de l'autre.

3) Chaque constituant décrit une partie d'une image

Lorsqu'ils se juxtaposent, les constituants créent alors une image métaphorique. Dans :

小橋流水 (xiǎoqiáo liúshuǐ) [petit pont, eau courante → un beau paysage]

on trouve les deux constituants « 小橋 » (petit pont) et « 流水 » (eau courante).

3.3.4 Mots composés de structure **AXBX**

Dans certains termes, les deux constituants XX s'intercalent avec deux formes A et B qui sont semblables, synonymes ou antonymes. Par exemple :

上**X**下**X** (shàng X xià X) [dessus X dessous X]　東**X**西**X** (dōng X xī X) [est X ouest X]

好**X**歹**X** (hǎo X dǎi X) [bon X mauvais X]　左**X**右**X** (zuǒ X yòu X) [gauche X droit X]

Trois cas de figure se présentent :

1) Les caractères X sont synonymes, par exemple :

左思右想 (zuǒ sī yòu xiǎng) [gauche réfléchir droite penser → réfléchir pendant longtemps]

2) Les caractères X appartiennent à la même classe sémantique, par exemple :

七手八腳 (qī shǒu bā jiǎo) [septs mains, huit pieds → agitation désordonnée]

Les mots simples « 手 » [main] et « 腳 » [pied] sont des membres du corps.

3) Les caractères X sont identiques, par exemple :

好說歹說 (hǎo shuō dǎi shuō) [bon parler mauvais parler → patiemment]

3.4 Locutions idiomatiques

On trouve beaucoup de locutions idiomatiques dans les textes chinois, et leur structure est plus ou moins régulière. Les locutions idiomatiques peuvent être formalisées avec NooJ de deux façons complémentaires : lorsque ces locutions sont insécables, on peut les ranger dans des dictionnaires, tout comme des mots composés ; lorsqu'elles admettent des insertions possibles (comme en français l'expression « prendre ... en compte »), on doit les traiter avec des grammaires syntaxiques locales, cf. (Silberztein 2007). Par exemple, dans l'expression suivante :

坐冷板凳 (zuò lěng bǎndèng) [s'asseoir sur un banc froid → être mal traité]

les constituants ne sont pas forcément juxtaposés, ce qu'on peut voir dans le texte suivant :

他已**坐**過了**冷板凳**。 (tā yǐ ***zuò*** guò le **lěng bǎndèng**) [Il s'est déjà trouvé dans la situation d'être mal traité]

Par ailleurs, certaines locutions sécables admettent des variantes orthographiques que l'on peut représenter dans les dictionnaires de NooJ. Par exemple, dans les deux expressions synonymes suivantes :

拖下水 (tuō xiàshuǐ) [implanter dans l'eau → impliquer quelqu'un dans une situation]

拉下水 (lā xiàshuǐ) [implanter dans l'eau → impliquer quelqu'un dans une situation]

扯下水 (chě xiàshuǐ) [implanter dans l'eau → impliquer quelqu'un dans une situation]

le verbe « 拖 » (tuō) peut être remplacé par les verbes « 拉 » (lā) et « 扯 » (chě) sans aucun changement de sens.

Certaines locutions admettent des variantes productives, que l'on peut traiter avec le module morphologique de NooJ, un peu comme on traite les variantes orthographiques des langues romanes. Par exemple, la locution suivante :

芝麻綠豆官 (zhīmá lǜdòu guān) [sésame, haricot mungo, fonctionnaire → petit fonctionnaire]

accepte quatre variantes que l'on peut représenter à l'aide d'un automate fini dans NooJ, cf. Figure 2.

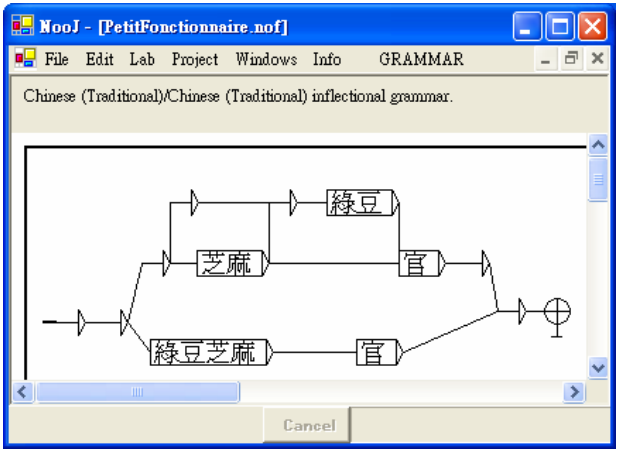


Figure 2 : Dérivation de la locution 芝麻綠豆官 [petit fonctionnaire]

4 Conclusion

Notre formalisation du vocabulaire chinois est fondée sur une classification en quatre classes d'unités linguistiques atomiques : affixes, mots simples, mots composés et expressions figées, qui correspondent à des critères purement orthographiques. Puisque la plupart des unités linguistiques chinoises sont indistinguables des simples séquences de caractères (il n'y a pas de blanc en chinois), il est indispensable de se doter de critères syntaxiques et sémantiques précis et reproductibles pour décider si une forme ou expression doit ou non être décrite dans un dictionnaire, plutôt que d'être traitée comme une séquence analysable de mots.

1) Le **critère de la compositionalité** vérifie si toutes les propriétés d'une forme peuvent ou non être calculées à partir des constituants de la forme.

2) Le **critère d'institutionnalisation** vérifie si une forme ou expression est ou non utilisée de façon systématique et privilégiée par rapport à d'autres expressions potentielles.

Les mots et expressions chinois sont souvent construits sur des schémas morpho-syntaxiques productifs (par exemple les polymères), bien connus des linguistes traditionnels chinois. Nous avons donc classé les termes selon leur structure, ce qui du même coup rend la maintenance de nos dictionnaires plus facile.

Nous avons dû modifier NooJ pour qu'il puisse prendre en compte des phénomènes spécifiques au chinois, comme par exemple la variation systématique de certains caractères (par exemple 羣=群) et les opérations morphologiques de réduplication (快樂 devient 快快樂樂). L'ensemble des outils lexicaux de NooJ (dictionnaires, grammaires flexionnelles, grammaires morphologiques et grammaires syntaxiques locales) ont été utilisés.

Le dictionnaire chinois de NooJ contient actuellement 93.013 entrées lexicales, dont 9 % sont des caractères autonomes (« mots simples ») ou non-autonomes, 27 % sont des mots de deux caractères, 24 % sont des mots composés de trois caractères et 40 % sont des mots de quatre

caractères ou plus. Le module chinois contient aussi des dictionnaires spécialisés, tels que un dictionnaire de noms de famille, un dictionnaire de toponymes, etc. ainsi que plus de vingt grammaires, dont des grammaires morphologiques et des grammaires syntaxiques locales, telles que la grammaire de date, de lieux, de noms de personne, etc.

Nous avons développé, testé et affiné ce module à partir d'un corpus d'une cinquantaine de textes littéraires (par exemple Sìshìtóngtáng de Lao She et *Chéngnán jiùshì* de Lin Haiyin) ; après enrichissement des données, le module couvre la totalité du vocabulaire de ces textes, en incluant la reconnaissance de motifs syntaxiques tels que les noms professionnels. Le module chinois de NooJ peut être téléchargé à partir du site WEB de NooJ : http://www.nooj4nlp.net.

Références

CAO WEI 曹炜. (2004). *Xiandai hanyu cihui yanjiu 现任汉语词汇研究*. Beijing 北京 : Peking University Press 北京大学出版社.

FU HUAIQING 符淮青. (2005). *Xiandai hanyu cihui (Zengding ben) 现代汉语词汇（增订本）*. coll. « yuyanxue jiaocai xilie 语言学教材系列 ». 2nd ed. Beijing 北京 : Beijing daxue chubanshe 北京大学出版社.

GE BENYI 葛本仪. (2002). *Xiandai hanyu cihuixue 现代汉语词汇学*. Jinan 济南 : Shandong renmin chubanshe 山东人民出版社.

KANG SHIYONG 亢世勇. (2004). *Mianxiang xinxi chuli de xiandai hanyu yufa yanjiu 面向信息处理的现代汉语语法研究*. coll. « Yuyan wenzi lilun yu yingyong yanjiu wenku 语言文字理论与应用研究文库 ». Shanghai 上海 : Shanghai cishu chubanshe 上海辞书出版社 and Shiji chuban jituan 世纪出版集团.

LIN HUEICHI 林惠祺. (2006). Les problèmes rencontrés dans le domaine de la catégorisation grammaticale du chinois et leurs solutions proposées. *9th INTEX/NooJ Conférence*.

LU SHICHENG 陸師成. (1992). *Cihai 辭海*. Taipei 臺北 : Wenhua tushu gongsi 文化圖書公司.

SILBERZTEIN MAX. (1993). Les groupes nominaux productifs et les noms composés lexicalisés. *Linguisticæ Investigationes* 2, 405-425.

SILBERZTEIN MAX. (2002). *NooJ Manuel* d'utilisation en anglais que l'on peut télécharger à partir de http://www.nooj4nlp.net (environ 200 pages).

SILBERZTEIN MAX. (2007). Frozen expressions and discontinuous annotations. In the Proceedings of Computational Linguistics 2007. Birmingham.

SUN YINXIN 孙银新. (2003). *Xiandai hanyu cisu yanjiu 现代汉语词素研究*. coll. « Zhonghua xueren congshu 中华学人丛书 ». Beijing 北京 : Zhongguo wenshi chubanshe 中国文史出版社.

ZHU DEXI 朱德熙. (2003). *Xiandai hanyu yufa yanjiu 现代汉语语法研究*. coll. « Shangwu yinshuguan wenku 商务印书馆文库 ». Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Étiquetage morpho-syntaxique de textes kabyles

Sinikka LOIKKANEN
Université d'Helsinki
sinikka.loikkanen@helsinki.fi

Résumé. Cet article présente la construction d'un étiqueteur morpho-syntaxique développé pour annoter un corpus de textes kabyles (1 million de mots). Au sein de notre projet, un étiqueteur morpho-syntaxique a été développé et implémenté. Ceci inclut un analyseur morphologique ainsi que l'ensemble de règles de désambiguïsation qui se basent sur l'approche supervisée à base de règles. Pour effectuer le marquage, un jeu d'étiquettes morpho-syntaxiques pour le kabyle est proposé. Les résultats préliminaires sont très encourageants. Nous obtenons un taux d'étiquetage réussi autour de 97 % des textes en prose.

Abstract. This paper describes the construction of a morpho-syntactic tagger developed to annotate our Kabyle text corpus (1 million words). Within our project, a part-of-speech tagger has been developed and implemented. That includes a morphological analyser and a set of disambiguation rules based on supervised rule-based tagging. To realise the annotation, a POS tagset for Kabyle is proposed. The first results of tests are very encouraging. At present stage, our tagger reaches 97 % of success.

Mots-clés : Étiquetage morpho-syntaxique, corpus de textes, langue kabyle, berbère.

Keywords: Part of speech tagging, text corpus, kabyle language, berber.

1 Introduction

L'étiquetage morpho-syntaxique automatique est une technologie relativement bien maîtrisée. Au moins pour les langues européennes comme l'anglais qui déjà dispose de grands volumes de corpus étiquetés. L'étiquetage morpho-syntaxique ou grammatical consiste à affecter à chaque occurrence d'un corpus un symbole représentant sa catégorie grammaticale (nom, verbe, ...) et, éventuellement, les informations morphologiques associées (masculin, singulier, ...) (Paroubek & Rajman, 2000). Les méthodes et les applications conçues pour une langue ne sont pas telles quelles transférables aux autres langues. Mais au niveau de l'annotation grammaticale, il y a eu d'importants efforts d'harmonisation et de standardisation aux seins de projets internationaux tels que MULTEXT (MUL, 1996) et EAGLES (EAG, 1996). Les recommandations et les standards de ces projets rendent possible la comparabilité des différents corpus étiquetés en plusieurs langues.

Dans le cadre de notre projet, nous avons constitué un corpus de textes kabyles, CKL (corpus kabyle littéraire). Le fond du corpus est constitué des six premiers romans kabyles publiés entre 1981–1995 dont le vocabulaire est présenté dans notre mémoire de DEA (Loikkanen, 1998). Par la suite, le corpus a été progressivement complété au fur et à mesure de la disponibilité

de nouveaux textes. Le CKL comprend actuellement soixante textes intégraux qui représentent un million d'occurrences. Les textes représentent différents genres littéraires (romans, contes, poèmes, chansons, récits) et couvrent différents thèmes. Les romans représentent la moitié des occurrences bien qu'ils ne représentent que 20 % des textes. Les textes sont numérisés, la structure balisée en XML[1] selon les principes de TEI[2] et la transcription est marquée par Unicode[3].

Le kabyle est une langue pour laquelle le travail d'étiquetage est une tâche nouvelle. Le kabyle est une variante de la langue berbère parlé en Kabylie, en Algérie du Nord. Le berbère ou *tamazight* (nom berbère de langue) est une langue afro-asiatique (ou chamito-sémitique) qui est considérée comme la langue autochtone de l'Afrique du Nord (Camps, 1987).

Pour annoter notre corpus, nous avons conçu et développé un étiqueteur ainsi qu'un jeu d'étiquettes morpho-syntaxiques. Dans ce qui suit, nous présentons d'abord notre jeu d'étiquettes morpho-syntaxiques pour le kabyle. Puis, nous présentons les principes de l'analyse. Pour finir, les premiers résultats de cette annotation seront rapportés.

2 Jeu d'étiquettes morpho-syntaxiques pour le kabyle

Pour annoter notre corpus, nous avons élaboré un jeu d'étiquettes morpho-syntaxiques pour le kabyle. Actuellement, il n'existe pas, à notre connaissance, d'études disponibles ni de catalogues avec un classement morpho-syntaxique fin pour le kabyle ou pour les autres dialectes du berbère. Dans les grammaires du kabyle (Mammeri, 1986; Mammeri, 1988; Naït-Zerrad, 2001), on trouve les descriptions pour les parties du discours comme nom, verbe, adjectif, adverbe, pronom, préposition, conjonction, etc. ; on a appliqué tel quel au berbère les catégories grammaticales que l'on retrouvait, par exemple, dans les grammaires traditionnelles du français. En fait, l'organisation des classes en berbère n'est pas radicalement différente de celle que l'ont peut rencontrer dans les langues indo-européennes (Chaker, 1984). Si l'on néglige un certain nombre de phénomènes secondaires et de formes isolées, les unités syntaxiques du berbère s'organisent en quatre grands ensembles qui sont 1) les verbes, 2) les noms, 3) les connecteurs ou relationnels et 4) les déterminants divers. Les deux premiers sont des catégories lexicales, les deux derniers des catégories grammaticales. Sous la classe de connecteurs, on regroupe tous les indicateurs de relation, prépositions, subordonnants, conjonctions et connecteurs divers. La classe des déterminants forme un ensemble hétérogène où on peut isoler, entre autres, une sous-catégorie comme les adverbes (Chaker, 1984).

Pour construire notre jeu d'étiquettes, nous avons défini un classement aussi fonctionnel que possible. Le jeu d'étiquettes s'inspire du projet EAGLES (EAG, 1996). Bien que ces recommandations ne soient pas applicables telles quelles au kabyle, elles servent de point de départ pour s'orienter vers des applications multilingues ainsi que la réutilisabilité et la comparabilité d'étiquettes. Pour l'instant, notre jeu d'étiquettes contient 12 catégories principales : nom (N), verbe (V), adjectif (A), pronom (PR), adverbe (ADV), préposition (PREP), conjonction (C), numéral (NU), interjection (I), particule (P), résidu (R) et ponctuation (PU). Ces catégories se basent sur les parties du discours et sur les traits morphologiques décrits dans les grammaires kabyles. Le jeu d'étiquetage est présenté par les paires attribut–valeur, par exemple un nom commun féminin singulier à l'état libre, comme *taqcict* 'une fille', est codé de la façon suivante :

[1] Extensible Markup Language, http ://www.w3.org/XML/
[2] Text Encoding Initiative, http ://www.tei-c.org/
[3] http ://unicode.org/

`N[type=commun genre=féminin nombre=singulier état=libre]` ou `NCFSL`.

– **Nom** : Le nom kabyle varie en genre, en nombre et en état[4]. Pour le nombre, outre les formes du singulier et du pluriel, il existe une forme duelle empruntée à l'arabe, mais très peu utilisée et on ne l'emploie que lorsqu'il s'agit de temps (*cehṛayen* 'deux mois'). L'état se manifeste par un changement d'une voyelle initiale et, éventuellement, par un ajout d'une semi-consonne (*w*, *y*) au début d'un mot pour les masculins commençant par une voyelle, par exemple *axxam* (état libre) 'maison' devient une forme *wexxam* (état d'annexion). Mais, il existe une partie importante de mots invariables en état (les noms de parenté, les emprunts au français et à l'arabe), i.e. les mots n'ayant pas de modifications dans la forme et dont l'état n'est pas ainsi identifiable automatiquement hors contexte. Pour le nom kabyle, nous avons ainsi défini les attibuts suivants dont les valeurs sont présentées entre crochets :

  type [commun, propre], genre [masculin, féminin], nombre [singulier, pluriel, duel],
  état [libre, annexion, non-marqué].

– **Verbe** : Les verbes kabyles sont flexionnels et varient en genre, en nombre et en personne. Il existe trois aspects : l'aoriste, l'aoriste intensif (inaccompli) et le prétérit (accompli), ainsi que trois modes : l'indicatif, l'impératif et le participe. L'infinitif n'existe pas au sens où on l'entend par exemple en français ; le lemme ou la forme lexicale d'un verbe donné dans les dictionnaires est la forme à l'impératif de la 2^{e} personne du singulier. Notre proposition pour les paires attribut–valeur pour les verbes kabyles est la suivante :

  mode [indicatif, impératif, participe], thème [aoriste, aoriste intensif, prétérit],
  degré [positif, négatif], nombre [singulier, pluriel], personne [1re, 2^{e}, 3^{e}],
  genre [masculin, féminin, commun].

– **Adjectif** : L'adjectif kabyle se forme principalement sur les verbes d'état (*aberkan* 'noir' du verbe *ibrik* 'être noir') ; par son type, il est qualificatif. Les formes secondaires sont les formes empruntées à l'arabe, les formes invariables et les formes complexes (Chaker, 1995). L'adjectif partage toutes les caractéristiques morphologiques du nom, il varie en genre, en nombre et en état, sauf les invariables. Les degrés de comparaison sont indiqués avec les verbes ou avec les prépositions, il n'y a pas de formes graphiques pour le comparatif ou pour le superlatif de type *bon – meilleur*. Les paires attribut–valeur proposées sont les suivantes :

  genre [masculin, féminin, commun], nombre [singulier, pluriel],
  état [libre, annexion, non-marqué].

– **Pronom** : Dans cette catégorie, nous avons regroupé différents types de pronoms et de déterminants, comme les pronoms personnels et les pronoms démonstratifs. Les pronoms personnels varient en genre, en nombre et en personne. Ils sont autonomes (*netta* 'il, lui') ou des affixes (*-k* 'à toi') liés à un élément comme nom, verbe, préposition, adverbe ou présentatif (l'affixe du nom exprime la possession, l'affixe du verbe marque le complément direct ou indirect, l'affixe des prépositions et l'affixe des adverbes joignent l'état ou l'action à une personne). Les démonstratifs forment un groupe déterminatif spécial. Il s'agit du groupe des unités fréquentes et largement cultivées dans le parlé indiquant la situation d'un objet par rapport au locuteur. Ils sont autonomes ou suffixés au noms ; les premiers varient en genre et en nombre, les derniers sont invariables. Les paires proposées sont les suivantes :

  type [personnel, démonstratif, indéfini, interrogatif, relatif],
  position [autonome, affixe], pers-type [nom, préposition, verbe],

[4]État : Il s'agit d'un concept grammatical représentant un couple oppositif état libre / état d'annexion dans lequel la forme du nom change. L'état d'annexion marque la dépendance du nom par rapport aux autres éléments de la phrase (Chaker, 1995).

nombre [singulier, pluriel], personne [1re, 2e, 3e], genre [masculin, féminin, commun], location [proximité, éloignement, absence].

- **Adverbe** : Mot invariable, sauf quelques exceptions qui portent la forme nominale et qui varient en état. Par son type, l'adverbe décrit le temps (*azekka* 'demain'), la manière (*baṭel* 'gratuitement'), la quantité (*mliḥ* 'beaucoup') ou le lieu (*beṛṛa* 'dehors').
- **Préposition** : Les prépositions sont invariables, utilisées isolément devant un nom (*zdat wexxam* 'devant la maison') ou avec les affixes personnels (*zdat-i* 'devant moi').
- **Conjonction** : Elles sont de types de coordination (*ihi* 'donc', *walakin* 'mais') et de subordination (*mi* 'quand', *qbel* 'avant que'). Certaines d'elles sont utilisées dans les deux cas.
- **Numéral** : Ils sont de types cardinal (*tlata* 'trois') ou ordinal (*wis tlata* 'troisième'). Ils varient en genre sauf les numéraux empruntés à l'arabe qui ne s'accordent pas en kabyle.
- **Particule** : Elles sont de petits mots invariables servant à préciser le sens d'autres mots. Dans les recommandations EAGLES, certains de ces éléments sont regroupés sous la catégorie 'unique/unassigned'. Les particules sont : particule de l'aoriste (*ad*), particule prédicative (*d*), particule de négation (*ur*), particule complétive de négation (*ara*), particule vocative (*a*), particule exclamative (*ack-*), particule présentative (*aql-*) et particule d'orientation (*d*, *n*).

3 Le problème de la variabilité des transcriptions

Les textes du CKL datent de différentes décennies, les premiers viennent de la fin du 19e siècle, les plus récents sont de l'année 2006. Bien que les principes de la notation (PNB, 1996; ALB, 1998) soient bien établis, on trouve dans la pratique des transcriptions différentes.

- Au niveau des lettres les variations se manifestent en différents choix graphiques :

  (1) *eçç* = *ečč* 'manger', *ḥemmel* = *ḫemmel* 'aimer', *ḍleb* = *ḏleb* 'demander', *imjuhad* = *imǧuhad* 'combattant', *taabbuṭ* = *taâbbuṭ* = *taɛbbuṭ* 'ventre'.

- Au niveau des mots, il y a des hésitations concernant le nombre des consonnes à écrire :

  (2) *claɣem* = *cclaɣem* 'moustache', *tagara* = *taggara* 'moment', *tidet* = *tidett* 'vérité',

  le placement d'un point emphatique :

  (3) *aḍaṛ* = *aḍar* = *adaṛ* 'pied', *aṣaḍuf* = *asaḍuf* 'loi', *lbaruḍ* = *lbaṛud* 'poudre à canon',

  ou la notation de la labio-vélarisation des consonnes vélaires et labiales :

  (4) *ameqran* = *ameᵠran* = *ameq°ran* = *ameqʷran* = *ameqwran* 'grand'.

- Au niveau des phrases, les variations se manifestent par l'assimilation phonétiques aux frontières des mots :

  (5) *ţ-ţideţ* ← *d tideţ*, *a d awi* ← *a d-tawi*.

  ou par la présence ou l'absence d'un tiret entre le nom ou le verbe et leurs affixes :

  (6) *yellis* = *yelli-s* = *yell-is* = *yelli s* 'sa fille'.

Finalement, la position du ə-muet pose certains problèmes. Il s'agit d'une voyelle neutre qui n'a pas de statut phonologique et dont la situation dépend de ce qui l'entoure. Par convention il est noté par une lettre 'e' dans tous les textes kabyles à quelques exceptions près. On ne peut pas nier son existence, au niveau graphique sa présence est importante : 12 % des lettres parmi toutes les lettres dans les six premiers romans kabyles (le deuxième rang après la lettre 'a'). Le placement de cet 'e' varie dû à l'enchaînement des mots :

(7) *iger* 'il a mis', mais *yegr-as* 'il lui a mis' au lieu d'écrire *iger-as*.

Pour gérer toutes ces variations et pour annoter le corpus, nous avons développé un analyseur qui se base sur les expressions régulières dans la reconnaissance des mots. Ainsi, nous trouvons des occurrences qu'elles soient écrites avec une ou plusieurs consonnes similaires au sein d'un mot (c{1,}laɣem), avec ou sans point sous la lettre (a(d|ḍ)a(r|ṛ)) ou qu'elles soient écrites différemment à cause de l'assimilation, comme pour le mot *abrid* 'chemin' à l'état d'annexion :

(8) [g] *brid* — [deǧǧ]-*ebrid* — [deg]-*gwebrid* — [deg]-*webrid* — [f]-*febrid* — [anebdu] *bwebrid* — [tikli] *bbwebrid* — *wubrid* — *ubrid.*

Le problème de la variabilité des transcriptions a ces conséquences aussi dans la segmentation des mots. La segmentation par des blancs n'est pas suffisante. Pour avoir des occurrences analogues, les unités textuelles sont ou découpées ou regroupées ou bien les deux à la fois :

– Les chaînes sont découpées lorsqu'il s'agit des affixes reliés par un trait d'union au mot auquel ils se rapportent. Parfois les mêmes unités sont écrites ensemble sans tiret entre les composants. Par exemple, les compositions *yemma s* | *yemma-s* | *yemmas* 'sa mère'.
– Les unités sont regroupées lorsqu'il s'agit des noms composées (*adrar ufud* 'tibia (montagne de jambe)', *Ait Ahmed*) ou les numéraux composés (*xems meyya* 'cinq cents'). Un grand groupe causant des hésitations dans l'écriture sont les prépositions complexes ; nous avons réuni les composants séparés par des blancs, par exemple, pour *seddaw* 'en-dessous de', nous avons aussi les formes *s ddaw* | *s eddaw* | *s-eddaw* | *si ddaw*.
– Parfois, les chaînes sont reconstruites. Par exemple, la composition comme *aqli-y-i* 'me voici' est découpée aux frontières des mots, puis, le reste a été regroupé pour avoir des occurrences analogues (*aql*|| *i-y-i*) à une forme « canoniques » *aql-iyi* 'me voici'.

4 Analyse morphologique

À cause de la transcription très instable, l'énumération de toutes les formes dans un lexique est une tâche impossible. En fait, nous n'avons pu utiliser les listes préalablement établies que pour les mots de classes grammaticales. Pour les autres classes, notamment pour les verbes et les noms, nous avons adopté une autre approche.

Bien que le kabyle soit une langue où la construction des mots se base principalement sur l'usage des racines, l'analyse se basant sur le modèle « interdigitation » de style sémitique n'est pas applicable telle quelle. Premièrement, parce que l'alternance vocalique n'est pas schématique ni toujours possible à prévoir, et deuxièmement, les racines kabyles ne portent pas un sens unique par racine. En fait, il existe plusieurs racines avec plusieurs champs sémantiques distincts, par exemple, la racine $\sqrt{br}$ a une vingtaine de sens différents.

Pour notre analyseur, nous avons fixé les exigences suivantes :

1. La langue que nous allons analyser se base sur l'utilisation des racines ; l'analyse doit produire de chaque forme au minimum un lemme et les consonnes de la racine.
2. L'analyseur doit identifier toutes les formes avec toutes les transcriptions possibles et rendre les descriptions morphologiques.
3. Pour identifier les formes, l'analyse se base sur l'extraction des stemmes, à partir desquels on dérive les lemmes et les racines.
4. Les dépendances de longues distances doivent être respectées, seulement les couples d'affixe « légaux » seront acceptés.

Par les couples d'affixe « légaux » on comprend les paires préfixes–suffixes correctes qui forment les mots. Par exemple, pour les verbes ordinaires nous avons 9 formes conjuguées en personne, 3 formes pour l'impératif et 5 formes de participe. Parmi ces formes, trois préfixes (∅-, `t-` et `i-`, dont le dernier peut se manifester aussi comme `y-`) et 3 suffixes (`-mt`, `-n` et -∅) sont ambigus. Les couples légaux sont les paires qui se réalisent correctement.

Pour construire un analyseur, on pourrait bien former trois listes dont la première pour les préfixes, la deuxième pour les stemmes et la troisième pour les suffixes, et construire un automate sur ces trois listes. Le problème dans cette approche est ce que le nombre de stemmes différents peut devenir relativement grand à cause des différentes notations utilisées dont nous ignorons a priori la nature des variations. Notre solution est l'identification des composants en expressions régulières. Initialement, nous n'avons les listes numériques que pour les affixes, pas celles de stemmes ou de racines. Ces dernières sont ici dérivées à partir des données du corpus. Notre analyseur identifie d'abord les composantes morphologiques des formes données, puis, à partir des stemmes obtenus, l'identification des racines et des lemmes est effectuée. Par exemple, pour la forme prétérit d'un verbe comme *turgamt* 'vous (*f.*) avez rêvé' la base de l'analyse est le stemme *urga* (racine : *rg* ; vocalisation : *u-a* ; stemme : *urga* 'prétérit' ; affixes : *t-, -mt* ; lemme : *argu* 'rêver') auquel on ajoute les affixes par la concaténation. Pour extraire ces composants, une règle est définie :

```
if ($w =~ /(^t)(.*?)mt$/) { $stemme = $2; }
```

Cette règle veut dire que si la forme commence par `t-` et finisse par `-mt`, la forme est vraisemblablement une forme verbale de la 2e personne féminine du pluriel. Ce qui reste, dans le milieu dans la variable $2, est le stemme avec toutes les transcriptions possibles. À partir de ce stemme nous obtenons les consonnes de la racine, le lemme et l'information concernant l'aspect verbal. Ce n'est que dans cette phase que l'on recourt à l'aide des dictionnaires pour définir la forme du lemme. Par exemple, pour les motifs *tegremt* et *tersemt*, pour lesquels on obtient les stemmes *egre* et *erse* respectivement, on défini ainsi avec les dictionnaires les formes du lemme qui sont pour ces verbes *ger* 'mettre' et *ers* 'descendre ; se poser ; se calmer'.

5 Désambiguïsation

On s'appuie sur l'approche à base de règles. On tente de minimiser l'ambiguïté par l'analyse du contexte proche ; ce processus repose sur l'hypothèse que la catégorie d'un mot dépend d'un contexte local, i.e. de la catégorie d'un mot précédent ou suivant (fenêtre = 1). Les définitions se basent ou sur les catégories grammaticales ou sur les catégories grammaticales et les traits morphologiques. Les règles contextuelles à définir sont de types :

```
« si X est marqué NOM et VERBE, et que X est précédé d'un Y marqué
PREP, alors l'étiquette VERBE doit être supprimée de X »
```

```
« si X = 'd', et que X est suivi d'un Z marqué NOM+EA, alors
X = 'PREP' »
```

Ces règles reposent sur l'analyse des mots proches qui peuvent eux-mêmes être ambigus, dans ces cas, les règles fonctionnent en cascade. Prenons par exemple la forme *yesli* (figure 1) dans une chaîne comme *axxam n yesli* 'la maison du jeune marié' (exemple 9) dans laquelle nous avons deux mots avec deux interprétations :

(9)

axxam	*n*	*yesli*
NOM	X	X
maison	?	?

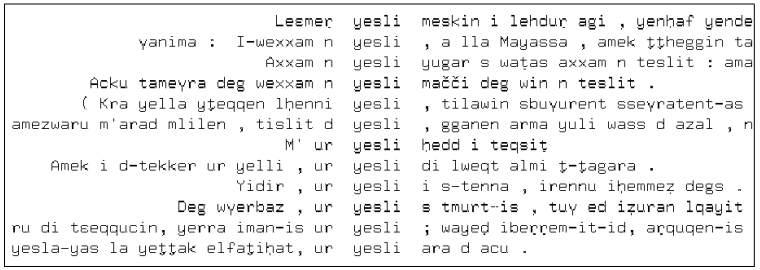


FIG. 1 – Début de la concordance de la forme *yesli* dans le CKL.

n : 1) préposition 'de, appartenant à' ; 2) particule d'orientation 'vers ici'.
yesli : 1) nom commun masculin singulier *isli* 'jeune marié', la forme à l'état d'annexion ; 2) prétérit négatif du verbe *sel* 'entendre', forme masculine 3e personne du singulier 'il n'a pas entendu'.

Les deux X sont résolus en deux phases : dans la première phase, on défini que *n* est une préposition s'il est précédé d'un nom ou d'un pronom et, dans la deuxième, on défini que *yesli* est un nom comme il a été précédé d'une préposition. Ces règles excluent les interprétations *n* = particule d'orientation et *yesli* = verbe qui se manifestent dans les phrases comme :

(10) *a n yuɣal* : particule de l'aoriste + particule d'orientation + verbe
ur yesli : particule de négation + verbe.

Dans certains cas, pour résoudre des ambiguïtés, la fenêtre a été étendue à couvrir les cas ± 2, et même plus, de manière à encadrer des classes grammaticales (affixes personnels, démonstratifs). Cela rend possible la résolution des cas comme :

(11) *axxam-nni n yesli* 'la maison (en question, dont on parle) du jeune marié'.

Les règles contextuelles se basant sur les catégories grammaticales ou sur les catégories grammaticales et les traits morphologiques ne peuvent pas toujours lever l'ambiguïté. Prenons comme exemple un mot kabyle très fréquent comme *d*, qui a, entre autres, les sens 1) particule prédicative 'c'est, ce sont' et 2) préposition 'et, avec'. Le sens de ce *d* peut être distingué, lorsqu'il précède un nom, par l'état de ce nom : si le nom qui suit est à l'état libre (EL), *d* a le sens 1 (exemple 12), si le nom qui suit est à l'état d'annexion (EA), *d* a le sens 2 (exemple 13).

(12) *argaz d aɣyul*
homme c'est âne+EL
l'homme est un âne

(13) *argaz d weɣyul*
homme et âne+EA
l'homme avec un âne

Mais, si le nom qui suit ne porte pas la marque de variation de l'état, comme par exemple dans l'énoncé *argaz d mmi-s* (exemple 14), nous avons deux sens selon le contexte : 1) l'homme, c'est son fils, 2) l'homme et son fils. Nous marquerons par EI (état invariable) les noms ne changeant pas de forme au début du mot.

(14) *argaz d mmi -s*
homme *d* fils+EI son
l'homme ? son fils

Dans ce dernier cas, pour enlever les ambiguïtés d'une façon automatique, on pensait à recourir aux méthodes statistiques de désambiguïsation, mais les premiers tests effectués ne confortent pas cette approche.

6 Tests et résultats

Nous avons testé notre système avec deux séries d'extraits de corpus. La première série a été utilisée comme corpus d'apprentissage pour la définitions des règles de désambiguïsation, la deuxième pour la validation. Les corpus de test ont été construits à partir des fragments de textes bruts de prose (contes, récits, romans). Pour la première série, on a pris de 6 textes différents les premiers 12000 bytes par texte. Cette quantité a puis été complétée pour terminer les dernières phrases obtenues jusqu'au point suivant terminant la phrase. Cela fait 1800–2500 mots par fragment (comptés par des blancs). Après la segmentation, le nombre se fixait entre 2000–2300.

Après l'analyse morphologique hors contexte, on constate qu'en moyenne un quart des mots était marqué comme ambigu. Sur la base de ces cas ambigus, on a écrit une centaine de règles de désambiguïsation. Les résultats sont vérifiés manuellement et, en cas d'erreurs, les règles sont corrigées, l'algorithme exécuté et le résultat vérifié. Finalement, on a obtenu 150 règles écrites. De cette manière on a construit itérativement un ensemble de règles de désambiguïsation. Après l'application de ces règles, il ne nous restait qu'en moyenne 1 % de cas ambigus. Le taux de réussite de la désambiguïsation est définit comme le pourcentage d'étiquettes assignées correctes par rapport au nombre total d'étiquettes assignées. Voir le tableau récapitulatif (tableau 1) où les résultats du premier tour avec les six fragments textuels.

1 N[a]	2 w[b]	3 <w>[c]	4 X S[d]	5 X S %[e]	6 X D[f]	7 %-final[g]
F1	1988	2314	623	26,9	31	98,6
F2	2514	2321	678	29,2	13	99,4
F3	2262	2189	488	22,3	23	98,9
F4	2064	2150	495	23,0	15	99,3
F5	2082	2052	382	18,6	8	99,6
F6	1803	2198	542	24,7	13	99,4

[a] Numéro de fragment.
[b] Nombre des mots w dans le texte brut, découpage par des blancs.
[c] Nombre des mots étiquetés en balise <w> ... </w> (hors ponctuation).
[d] Nombre des cas X, i.e. des mots ayant de multiples interprétations, dans les textes segmentés correctement après l'analyse morphologique, avant l'application des règles de désambiguïsation.
[e] Précédant en pourcentage.
[f] Nombre des cas X restant ambigus après l'application des règles de désambiguïsation.
[g] Pourcentage d'étiquettes réussites.

TAB. 1 – Résultats du premier tour.

Les cas restants encore ambigus, environ 1 %, sont les cas où $X = d$ et pour lesquels nous n'avons pas pu écrire des règles, par exemple, comme dans la phrase suivante (exemple 15) :

(15) *yekseb* *tameṭṭut* ***d*** *lall* *n* *nnif* ***d*** *lḥeṛma*
VPS3M NCFSL **X** NCFSI PREP NCMSI **X** NCFSI
il-a-possédé une-femme **?** propriétaire de honneur **?** respect
il avait une femme (**c'est** une propriétaire) d'honneur **et** de sacrée

Pour désambiguïser ces cas automatiquement, on a examiné les distributions des cas $X = d$ dans les fragments du corpus d'apprentissage en gardant à l'esprit les méthodes probabilistes de désambiguïsation. Les nombres obtenus étaient

$d = PO$ (particule d'orientation) : 50 %
$d = PD$ (particule prédicative) : 40 %
$d = PREP$ (préposition) : 10 %.

Parmi les cas qui restaient encore ambigus et que nous avons ensuite annotés manuellement, la distribution des cas $X = d$ est la suivante :

$d = PO$ (particule d'orientation) : 0 %
$d = PD$ (particule prédicative) : 60 %
$d = PREP$ (préposition) : 40 %.

Ces nombres indiquent que pour résoudre les derniers cas $X = d$ (+ N+EI, nom à l'état invariable) la résolution doit être cherchée ailleurs ; les méthodes probabilistes simples ne sont pas capables de les résoudre lorsqu'on obtient des répartitions presque égales (40 % / 60 %).

La deuxième série de test (tableau 2), validation des règles, a été effectuée avec deux fragments de texte plus larges. Pour cela, on a pris les premiers 50000 bytes de deux textes bruts en prose (un roman, un conte) qui ont ensuite été complétés pour terminer les dernières phrases obtenues jusqu'au point suivant terminant la phrase. Ces deux textes n'ont pas été utilisés dans la dérivation des règles de désambiguïsation (serie 1 ci-dessus). La validation est faite avec les règles définies dans le premier test. Cette fois, le taux de résolution du texte a atteint 97 %. En examinant les listes des cas restant non résolus, on constate qu'il y a 1) des cas où $X = d$ (+ N+EI), 2) des nouveaux cas X pour lesquels il n'y avait pas encore de règles ainsi que 3) des cas pour lesquels les règles déjà existaient mais qui restaient ambigus à cause de la transcription variable.

1 N[a]	2 w[b]	3 <w>[c]	4 X S[d]	5 X S %[e]	6 X D[f]	7 %-final[g]
T1	8289	9335	2169	23,2	162	98,2
T2	8074	9007	2055	22,8	239	97,3

[a−−g] Voir les explications ci-dessus, tableau 1.

TAB. 2 – Résultats du deuxième tour.

7 Conclusions et perspectives

Dans cet article, nous avons présenté notre travail sur l'étiquetage d'un corpus kabyle littéraire. L'étiquetage morpho-syntaxique est une tâche nouvelle dans le domaine du berbère. Il n'existe, à notre connaissance, aucun corpus kabyle, ni berbère, sous format numérique qui soit grammaticalement annoté, publié et disponible. Au sein de notre projet, nous avons construit

un étiqueteur, un jeu d'étiquettes morpho-syntaxiques ainsi qu'un formalisme pour résoudre les ambiguïtés dues aux homographes. Notre étiqueteur est constitué de trois modules : le segmenteur, l'analyseur et le désambiguïsateur. Dans toutes les étapes, les spécificités du corpus ont été prises en compte.

Les résultats préliminaires des tests effectués sont très encourageants. La désambiguïsation faite à base de règles a pu atteindre un taux d'étiquetage correct de 97 % des textes en prose.

La construction d'un étiqueteur morpho-syntaxique pour une langue pour laquelle les ressources numériques (lexiques, textes annotés) et les définitions formelles sont encore en cours de développement est un processus qui avance pas à pas. Avec notre méthode nous produisons à partir des données du corpus les ressources qui manquent : les premières listes numériques des stemmes, des lemmes et des racines qui peuvent être utilisées ultérieurement dans les futures versions de l'étiqueteur. Dans les perspectives d'évaluation du jeu d'étiquettes, une analyse plus profonde concernant les parties du discours et la catégorisation des mots grammaticaux est envisagée. La précision et l'extension des règles de désambiguïsation ainsi que de nouveaux tests se basant sur ces règles plus étendues avec des données sur une grande échelle sont aussi prévues.

Références

ALB (1998). *Aménagement linguistique de la langue berbère*. Inalco, Paris. `http://www.inalco.fr/crb/docs_pdf/amenage1998.pdf` [10.9.2006].

CAMPS G. (1987). *Les Berbères. Mémoire et identité*. Paris : Éditions Errance.

CHAKER S. (1984). *Textes en Linguistiques Berbère. Introduction au domaine berbère*. CNRS.

CHAKER S. (1995). *Linguistique berbère. Étude de syntaxe et de diachronie*. Paris : Peeters.

EAG (1996). EAGLES, Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R. `http://www.ilc.cnr.it/EAGLES96/home.html` [15.1.2007].

LOIKKANEN S. (1998). Vocabulaire du roman kabyle (1981–1995). Une étude quantitative. *Études et documents berbère*, **15–16**, 185–196.

MAMMERI M. (1986). *Précis de grammaire berbère (kabyle)*. Paris : AWAL.

MAMMERI M. (1988). *Tajeṛṛumt n tmaziɣt (tantala taqbaylit)*. Paris : AWAL/La Découverte.

MUL (1996). Multilingual Text Tools and Corpora. `http://www.lpl.univ-aix.fr/projects/multext` [15.1.2007].

NAÏT-ZERRAD K. (2001). *Grammaire moderne du kabyle. Tajerrumt tatrart n teqbaylit*. Paris : Karthala.

PAROUBEK P. & RAJMAN M. (2000). Étiquetage morpho-syntaxique. In J.-M. PIERREL, Ed., *Ingénierie des langues*, p. 131–150. Paris : HERMES Science Europe.

PNB (1996). *Propositions pour la notation usuelle à base latine du berbère*. Inalco, Paris. `http://www.inalco.fr/crb/docs_pdf/notation.pdf` [10.9.2006].

Analyse syntaxique et traitement automatique du syntagme nominal grec moderne

Athina MICHOU
LATL – Université de Genève
Athina.Michou@lettres.unige.ch

Résumé. Cet article décrit le traitement automatique du syntagme nominal en grec moderne par le modèle d'analyse syntaxique multilingue Fips. L'analyse syntaxique linguistique est focalisée sur les points principaux du DP grec : l'accord entre les constituants fléchis, l'ordre flexible des constituants, la cliticisation sur les noms et le phénomène de la polydéfinitude. Il est montré comment ces phénomènes sont traités et implémentés dans le cadre de l'analyseur syntaxique FipsGreek, qui met en œuvre un formalisme inspiré de la grammaire générative chomskyenne.

Abstract. This article describes an automatic treatment to the Modern Greek noun phrase in terms of the Fips multilingual syntactic parser. The syntactic analysis focuses on the main issues related to the Greek DP: the agreement among the inflected constituents, the relatively free constituent order, noun cliticisation, and the polydefiniteness phenomenon. The paper discusses how these processes are treated and implemented within the FipsGreek parser, which puts forth a formalism relying on Chomsky's generative grammar.

Mots-clés : analyseur grec, analyse morphosyntaxique, syntagme nominal, grec moderne.

Keywords: Greek parser, morphosyntactic analysis, determiner phrase, modern Greek.

1 Introduction

C'est aux années '70 que remonte l'étude du grec moderne sous l'angle des théories linguistiques modernes; c'est la période que marquent la reconnaissance et la propagation de la grammaire générative aux Etats-Unis et en Europe, en tant que modèle dominant de la recherche dans le domaine syntaxique des langues naturelles. Depuis, les études de la morphosyntaxe du grec moderne ont connu une évolution croissante jusqu'à nos jours où la recherche linguistique constitue un champ bien établi.

Parmi d'autres champs de la syntaxe grecque, le syntagme nominal et ses aspects ont souvent offert des sujets de débats et d'analyses linguistiques étendues. En particulier, l'attention des chercheurs a été attirée par la place et l'ordre relativement libre des modificateurs nominaux, le comportement des noms déverbaux, la nominalisation, les pronoms déictiques, la distribution et l'interprétation du génitif possessif. En outre, un défi pour les chercheurs a constitué le phénomène de la répétition de l'article défini accompagné d'un adjectif, appelé construction polydéfinie, détermination multiple, double définitude ou encore *determiner*

spreading. Néanmoins, si la situation au niveau de la recherche linguistique est florissante, le domaine du traitement automatique de la syntaxe grecque n'est pas équivalent. Cela est du au fait de la complexité de la langue, vu l'ordre relativement libre de ses constituants, et au manque des ressources linguistiques pour l'évaluation des travaux. Ainsi, la plupart de programmes de traitement du langage développés jusqu'à aujourd'hui visent à traiter la morphologie mais très peu la syntaxe du grec moderne. De plus, les analyseurs morphosyntaxiques intégrés dans les processeurs lexicaux des systèmes du traitement automatique du grec sont d'habitude destinés à la correction orthographique ou syntaxique ou encore à une application de lexicographie ou de la recherche de l'information.

Dans le cadre de la présente recherche, nous avons envisagé d'étudier les phénomènes intéressants du syntagme nominal en grec moderne dans le but de bâtir un analyseur syntaxique qui puisse traiter efficacement sa structure. La structure du syntagme nominal grec présente des particularités que les langues romanes et germaniques ne partagent pas et dont les analyseurs syntaxiques existants ne tiennent pas compte. Dans notre étude, nous proposons une analyse linguistique de la structure du syntagme nominal grec en vue de son traitement automatique par le modèle FipsGreek (Leoni de León & Michou, 2006). Du point de vue théorique, notre approche repose sur une analyse profonde basée sur l'hypothèse dite analyse DP et du point de vue computationnel la modélisation prend place dans le cadre d'analyse automatique multilingue.

2 État de l'art

Pendant les deux dernières décennies, plusieurs tentatives de traiter le grec moderne, notamment au niveau du traitement morphologique et du traitement syntaxique ont vu le jour dans des instituts et départements universitaires du pays. Parmi ces travaux, il y en a qui traitent des données linguistiques spécifiques sans le support ferme d'un cadre théorique, d'autres qui se réfèrent à des applications au grec moderne des outils puissants déjà développés, comme par exemple le modèle KIMMO (Markopoulos, 2001) et encore d'autres qui recouvrent diverses approches linguistiques théoriques dans le TALN tels que Affix Grammars (Triantopoulou, 1995), HPSG (Kolliakou, 2004), grammaire de validation (Ralli & Galiotou, 1987). Néanmoins, la plupart d'entre eux, même s'ils ont donné des résultats satisfaisants, réalisent un traitement morphologique ou syntaxique partiel. Parmi les programmes développés, nous allons citer ceux qui ont pour but de traiter la syntaxe du grec moderne.

Triantopoulou (1995) a présenté une première tentative du traitement automatique du grec moderne. Son travail consiste en une description du syntagme nominal suivant le modèle formel des grammaires d'affixes AGFL (Affix Grammar over a Finite Lattice) et en utilisant un système orienté vers les corpus.

Draggiotis, Grigoriadou et Philokyprou (1997) ont présenté le système Dinous, dans le but de simuler la manière dont les locuteurs natifs analysent les phrases naturelles. Cet analyseur combine deux sous-systèmes computationnels différemment structurés. Le premier d'entre eux utilise de l'information concernant les préférences des locuteurs natifs et le deuxième traite la connaissance linguistique. Dinous analyse des phrases en fournissant comme résultats des représentations arborescentes annotées avec de l'information morphosyntaxique.

Kermanidis, Sgarbas, Fakotakis et Kokkinakis (2000) ont implémenté un analyseur syntaxique pour le grec moderne basé sur le formalisme PC-PATR, ayant pour objectif de traiter plusieurs phénomènes syntaxiques. Dans ce cadre, ils traitent les constructions à verbes copulatifs, l'ensemble des propositions subordonnées, tous les types d'accord (accord entre déterminant - adjectif et nom, sujet - verbe) et la majorité des possibilités de l'ordre des mots

dans les structures syntagmatiques.
Baldzis, Eumeridou et Kolalas (2002) ont développé un système pour le traitement automatique du grec moderne, inscrit dans le domaine de l'enseignement des langues assisté par ordinateur (ELAO). Le modèle proposé essaie, d'une part de formaliser la morphologie, la syntaxe et la sémantique du grec moderne, et d'autre part d'encoder dans un ensemble fini des règles les interactions entre ces niveaux, au cours de l'analyse syntaxique des propositions réelles. Le système a été créé pour couvrir principalement des domaines thématiques importants pour l'approche communicative de l'enseignement et peut être utilisé dans la classe ou par correspondance via Internet.

3 Description du syntagme nominal grec moderne

Dans cette partie, nous allons nous rapporter aux constituants divers du syntagme nominal, et aborder des phénomènes intéressants qui s'y produisent : l'accord entre les constituants fléchis et leur ordre (section 3.1), la cliticisation (section 3.2), la polydéfinitude (section 3.3).

3.1 Les propriétés des constituants du syntagme nominal

Le syntagme nominal du grec moderne est composé d'une tête que constitue le nom, de modificateurs possibles que sont le déterminant, l'adjectif, l'adverbe, et de compléments tels qu'un syntagme nominal au génitif, un pronom clitique ou une proposition relative. Les noms, ainsi que tous ses modificateurs fléchis (déterminants, numéraux, adjectifs) sont dotés d'un genre (masculin - féminin - neutre), d'un nombre (singulier - pluriel), d'un cas (nominatif - accusatif - génitif - vocatif) et appartiennent à plusieurs classes flexionnelles qui déterminent leurs suffixes flexionnels aux cas du singulier et du pluriel. Les caractéristiques majeures des constituants fléchis du syntagme nominal sont : l'accord entre eux en genre, nombre et cas, comme le dénote clairement l'exemple (1a-f), et leur ordre relativement libre :

(1) a. *Ο μικρός πρίγκιπας* (nom-masc-sg)
b. *Του μικρού πρίγκιπα* (gen-masc-sg)
c. *Το μικρό πρίγκιπα* (acc-masc-sg)
Det_{def} $\text{petit}_{\text{adj}}$ $\text{prince}_{\text{Nom}}$
'Le petit prince'

d. *Οι μικροί πρίγκιπες* (nom-masc-pl)
e. *Των μικρών πριγκίπων* (gen-masc-pl)
f. *Τους μικρούς πρίγκιπες* (acc-masc-pl)
Det_{def} $\text{petits}_{\text{adj}}$ $\text{princes}_{\text{Nom}}$
'Les petits princes'

Les deux articles, l'article défini (*ο, η, το* 'le, la') et l'article indéfini (*ένας, μια, ένα* 'un, une'), marquent la définitude ou l'indéfinitude d'un syntagme nominal. Il se peut, cependant, que l'article soit absent. En ce qui concerne l'ordre des éléments divers du syntagme nominal, les modificateurs précédent majoritairement le nom en se plaçant normalement entre l'article et le nom. Un autre modificateur ou un adjectif uniquement peut les séparer par la tête nominale, à l'exception du démonstratif *τέτοιος* 'tel' et du contrastif *άλλος* 'autre' qui peuvent soit précéder soit suivre les noms. Les démonstratifs *αυτός* 'ce', *τούτος* 'ce-ci' et *εκείνος* 'ce-là', ainsi que les quantifieurs *όλος* 'tout' et *ολόκληρος* 'entier' diffèrent des autres modificateurs en restant à la périphérie du syntagme nominal. Dans le cas où plusieurs modificateurs sont employés ensemble dans le syntagme nominal, leur ordre canonique est le suivant : *Quantifieur + démonstratif + déterminant + numéral + adjectif + nom + pronom clitique* (Holton et al., 1997).

3.2 La cliticisation dans le syntagme nominal

Les deux types de complément du nom dont nous tenons compte dans cette étude sont : I. Le syntagme nominal plein au génitif, en l'occurrence le syntagme *της Ειρήνης*, exprimant le possesseur qui peut soit suivre (ex. 2a), soit précéder le nom (ex. 2b) :

(2) a. *Το αυτοκίνητο της Ειρήνης*
Det_{def} $voiture_{Nom}$ Det_{def} $Irène_{Nom}$
'La voiture d'Irène'

b. *Της Ειρήνης το αυτοκίνητο*
Det_{def} $Irène_{Nom}$ Det_{def} $voiture_{Nom}$
'La voiture d'Irène'

II. Le génitif des pronoms clitiques (*μου, σου, του/της, μας, σας, τους*) qui, en tant qu'élément possessif, s'attache toujours à un élément hôte qui peut être soit la tête nominale (ex. 3a, 4a), soit un numéral (ex. 3b), soit un adjectif (ex. 4b), mais jamais l'article défini (ex. 3c-d, 4c-d). D'autre part, le clitique *της* ne peut pas apparaître seul sans être soudé à un élément hôte (ex. 3e). A savoir, il ne se comporte ni comme un déterminant, ni comme un adjectif ; il ne partage pas les propriétés distributionnelles de ces catégories et il dispose des propriétés particulières (étant intransitif, ne faisant pas l'accord avec son hôte, portant des traits de personne) et il entretient un lien thématique avec le nom dont ils est sélectionné.

(3) a. *Το αυτοκίνητό της*
Det_{def} $voiture_{Nom}$ $PRcl_{gén}$
'Sa voiture'

b. *Το πρώτο της αυτοκίνητο*
Det_{def} $premier_{Num}$ $PRcl_{gén}$ $voiture_{Nom}$
'Sa première voiture'

c.**Το της αυτοκίνητο*

d.**Το της πρώτο αυτοκίνητο*

e.**Της αυτοκίνητο*

(4) a. *Το ωραίο αυτοκίνητό της*
Det_{def} $belle_{adj}$ $voiture_{Nom}$ $PRcl_{gén}$
'Sa belle voiture'

b. *Το ωραίο της αυτοκίνητο*
Det_{def} $belle_{adj}$ $PRcl_{gén}$ $voiture_{Nom}$
'Sa belle voiture'

c.**Το της ωραίο αυτοκίνητο*

d. **Το της το ωραίο το αυτοκίνητο*

3.3 Les adjectifs et le phénomène de la polydéfinitude

En général, les modificateurs adjectivaux peuvent apparaître en position prénominale ou postnominale. Plus précisément, les adjectifs dans leur usage attributif sont prénominaux étant précédés de l'article défini (ex. 5a) ou indéfini (ex. 6a). Ils peuvent également suivre le nom dans un syntagme nominal indéfini (ex. 6b), mais jamais dans un syntagme nominal défini simple (ex. 5b). La répétition de l'article défini est obligatoire devant l'adjectif qui apparaît en position postnominale seulement si la tête nominale est définie (ex. 5c); au contraire, la propagation de l'article indéfini est exclue (ex. 6c):

(5) a. *Η* ***πολυτελής*** *κατοικία*
Det_{def} $luxueux_{adj}$ $domicile_{Nom}$

b.**Η κατοικία* ***πολυτελής***
Det_{def} $domicile_{Nom}$ $luxueux_{adj}$
'Le luxueux domicile'

c. *Η κατοικία η* ***πολυτελής***

(6) a. *Μια* ***πολυτελής*** *κατοικία*
Det_{indef} $luxueux_{adj}$ $domicile_{Nom}$

b. *Μια κατοικία* ***πολυτελής***
Det_{indef} $domicile_{Nom}$ $luxueux_{adj}$
'Un domicile luxueux'

c. **Μια κατοικία μια* ***πολυτελής***

Cette possibilité d'avoir à l'intérieur du syntagme nominal grec des occurrences multiples du déterminant défini suivi par des adjectifs, constitue une particularité appelée 'determiner spreading' (Androutsopoulou, 1995) ou 'construction polydéfinie' (Kolliakou, 2004). Ce cas de figure est mieux illustré dans les exemples (7a-c), où l'article défini *το* 'le' précède obligatoirement le nom *αυτοκίνητο* 'voiture' ainsi que les adjectifs *μικρό* 'petit' et *ιταλικό*

'italien' qui le modifient. Les adjectifs qui entrent dans la polydéfinitude peuvent être également postnominaux (ex. 7a), prénominaux (ex. 7b) ou les deux (ex. 7c-d). Par contre, le syntagme nominal en (ex. 7e) est agrammatical puisqu'il n'existe pas d'article défini pré-adjectival. En outre, le placement des adjectifs dans les exemples ci-dessous est interchangeable en rendant tous les différents ordres possibles :

(7) a. *Το αυτοκίνητο το μικρό το ιταλικό*
Det_{def} $voiture_{Nom}$ Det_{def} $petite_{adj}$ Det_{def} $italienne_{adj}$
'La petite voiture italienne'
b. *Το μικρό το ιταλικό το αυτοκίνητο*
c. *Το μικρό το αυτοκίνητο το ιταλικό*
Det_{def} $petite_{adj}$ Det_{def} $voiture_{Nom}$ Det_{def} $italienne_{adj}$
d. *Το ιταλικό το αυτοκίνητο το μικρό*
e. **Το αυτοκίνητο μικρό ιταλικό*

Comme il sera présenté dans la section 4.1, dans notre analyse nous adoptons le cadre théorique de l'analyse DP (Abney, 1987) qui postule que le syntagme nominal est la projection d'une catégorie fonctionnelle D qui est réalisée par un élément appelé déterminant et sélectionne un complément de catégorie NP. Pour ce qui est de l'attachement des adjectifs, lorsqu'ils sont prénominaux, ils constituent des spécificateurs du nom et ils occupent la place de [Spec, NP] (Stavrou, 1996, 1999). En plus, la position canonique des adjectifs étant à gauche de la tête nominale, l'accord morphosyntaxique réalisé entre l'adjectif et le nom est le type d'accord qui a lieu directement entre Spec - tête nominale. Quant aux adjectifs postnominaux, ils ont une lecture prédicative et ils sont analysés dans une projection fonctionnelle (FP) de type small clause, adjointe à droite du NP dans laquelle le nom n'est pas réalisé. La procédure de l'analyse Fips de ce type de structure sera décrite plus en détail en fin de la section 4.2.

Dans notre analyse, nous traitons aussi les séquences d'adjectifs, en usage attributif, qui sont soit en modification parallèle du nom, c'est-à-dire une suite linéaire d'adjectifs, sans marque de ponctuation ou de conjonction (ex. 8a), soit en modification avec marque de conjonctions (ex. 8b). Dans la structure en (8b) les adjectifs étant coordonnés par la conjonction de coordination *και* 'et', ils constituent des conjoints dans une projection ConjP. En ce qui concerne l'ordre d'apparition des adjectifs dans la projection nominale incluse dans les structures suivantes, il est flexible permettant toutes les combinaisons possibles[1], à l'exception du quantifieur adjectival *πολλά* 'beaucoup de', qui doit précéder les autres adjectifs.

(8) a. *Τα πολλά μικρά εικονογραφημένα παιδικά βιβλία* ⟶ [DP [NP [AP][AP][AP][AP] N]]
Det_{def} $nombreux_{adj}$ $petit_{adj}$ $illustré_{adj}$ $enfant_{adj}$ $livre_{Nom}$
'Les nombreux petits illustrés livres pour enfants'
b. *Ένα ωραίο και γρήγορο αυτοκίνητο* ⟶ [DP[NP[AP [ConjP[AP] [AP]]] N]]
Det_{indef} $belle_{adj}$ et $rapide_{adj}$ $voiture_{Nom}$
'Une belle et rapide voiture'

[1] Sachant qu'au point de vue linguistique il y a des restrictions dans l'ordre des adjectifs qui semblent être des préférences, certaines combinaisons apparaissent plus ou moins acceptables au niveau du parsing.

4 Analyse du groupe nominal proposée dans Fips

4.1 Le formalisme du système Fips et l'algorithme d'analyse

Dans notre système FipsGreek, tout comme dans les autres Fips faisant partie du modèle d'analyse syntaxique multilingue (Wehrli, 2004), le formalisme grammatical utilisé relève de la théorie générative chomskyenne et notamment du module X-barre qui dicte de façon générale et uniforme la géométrie des structures syntaxiques. Néanmoins, le schéma X-barre est simplifié pour faciliter le processus d'implémentation. Comme le nœud 'barre' n'est pas représenté au niveau informatique le schéma est transformé à celui présenté dans la figure 1. Ainsi, la tête syntagmatique X, est une catégorie lexicale (N/nom, V/verbe, P/préposition, Adv/adverbe, Adj/adjectif, Conj/conjonction) ou fonctionnelle (C/complementeur, T/verbe conjugué, D/déterminant, F/fonctionnelle), ainsi que Left (à gauche de la tête) et Right (à droite de la tête) correspondent à des listes (éventuellement vides) de projections maximales.

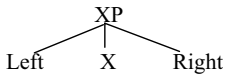


Figure 1 : le schéma X-barre simplifié

Par convention, dans le schéma, la projection DP (Determiner Phrase) est la représentation d'un syntagme dont la tête D (Abney, 1987), réalisée par un déterminant, sélectionne une catégorie syntaxique NP en position de sous-constituant droit. En ce qui concerne l'algorithme d'analyse, il comprend une analyse lexicale, responsable de la segmentation d'une phrase d'entrée en unités lexicales, servant de base à une analyse syntaxique, qui assigne à une phrase une ou plusieurs structures syntaxiques. Puisque la grammaire n'est pas déterministe, c'est-à-dire qu'elle admet plusieurs alternatives possibles, l'analyseur doit traiter plusieurs hypothèses possibles en parallèle. Cela se fait par une stratégie d'analyse syntaxique de type gauche-droite avec un traitement parallèle des alternatives, combinant une approche incrémentale, essentiellement ascendante avec un filtre descendant. En d'autres termes, pour chaque nouveau mot lu, l'analyseur cherche à le combiner avec les constituants déjà construits (contexte gauche), au moyen des opérations suivantes : i) Projection : un constituant maximal (XP) est projeté sur la base des traits inhérents du mot lu. ii) Attachement à gauche : le constituant étant à gauche du nouveau constituant est attaché comme sous-constituant du nouveau constituant. iii) Attachement à droite : le nouveau constituant gauche est attaché comme complément d'un constituant du contexte gauche.

4.2 Les règles proposées et le traitement des données

Dans cette section, nous présenterons notre proposition d'intégrer dans l'analyseur FipsGreek les structures du syntagme nominal grec que nous avons déjà présentées dans la section 3. Comme nous venons le voir, dans le formalisme du système Fips les constituants syntagmatiques sont attachés à gauche et à droite des catégories syntaxiques. En ce qui concerne l'analyse du syntagme nominal, nous adoptons l'hypothèse théorique portant sur l'analyse du DP (Abney, 1987) selon laquelle le syntagme nominal se compose de la projection d'une catégorie fonctionnelle D, remplie généralement par un déterminant lexical, et qui prend comme complément une projection lexicale de catégorie NP (Horrocks & Stavrou, 1987). Ainsi, nous obtenons le syntagme $[_{DP}$ [D $[_{NP}$ [N]]]], illustré dans la figure

2(a), où la catégorie D peut être occupée par une catégorie lexicale autre que l'article, comme illustré dans 2(b), et elle peut sélectionner un complément NP ou un autre D (cf. 2c), d'après les règles de sélection déterminées dans notre grammaire :

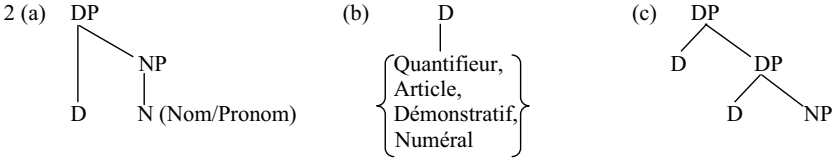


Figure 2(a-c): la structure du DP grec

En grec, comme en français, les éléments appartenant à la catégorie D sélectionnent un NP dont la tête doit être lexicalement réalisée. Par exemple, la structure du DP *το αυτοκίνητο* 'la voiture' est rendue ainsi : [$_{DP}$ *το* [$_{NP}$ *αυτοκίνητο*]]. Dans ce syntagme, la tête nominale *αυτοκίνητο*, occupe la position N et sa projection maximale (NP) s'attache comme complément de D. En plus, l'analyseur, pendant le processus de la reconnaissance des mots, doit vérifier l'accord en genre, nombre et cas, réalisé entre le D et le NP (accord tête-compl) ; ce type d'accord est obligatoire entre tous les déterminants fléchis (articles, quantifieurs, numéraux) et le nom et fait partie intégrale des règles d'attachement des constituants. Dans notre analyse, chaque déterminant est transposé en tant que tête D sélectionnant soit un autre DP (ex. 9) soit un NP comme complément. En outre, dans le lexique chaque entrée lexicale est spécifiée pour tous les traits morphologiques, sélectionnels et sémantiques nécessaires afin d'être utilisés au cours de l'analyse pour autoriser les différents types d'attachements. Ainsi, une séquence de différents types de déterminants, comme celle de l'exemple (9), sera analysée sur la base des traits de sélection qui sont associés à chaque lexème. La structure de sortie du syntagme nominal en (9) est présentée en (10) :

(9) *Όλες αυτές οι δέκα ετήσιες πολιτιστικές εκδηλώσεις του συλλόγου μας*
Det_{quant} Det_{dem} Det_{def} Det_{num} Adj Adj activité Det_{def} association $PRcl_{gén}$
'Toutes ces dix annuelles activités culturelles de notre association'

(10) [$_{DP}$ *όλες* [$_{DP}$ *αυτές* [$_{DP}$ *οι* [$_{DP}$ *δέκα* [$_{NP}$ [$_{AP}$ *ετήσιες*] [$_{AP}$ *πολιτιστικές*] *εκδηλώσεις*[$_{DP}$ *του* [$_{NP}$ *συλλόγου* [$_{N(+CL)}$ *μας*]]]]]]]]

Or, quelques déterminants ont un comportement distributionnel plus compliqué que d'autres. Cette catégorie contient les quantifieurs *όλος* 'tout' et *ολόκληρος* 'entier', ainsi que les démonstratifs *αυτός* 'celui-ci', *εκείνος* 'celui-là' et *τούτος* 'ce-ci' qui sont prénominaux en tant que déterminants, ou postnominaux, ayant un comportement adjectival (cf. *Όλος ο κόσμος* et **ο όλος κόσμος* 'tout le monde' vs. *Ο κόσμος όλος* 'le monde entier'). Pour traiter ces éléments, nous avons opté pour une double entrée lexicale : celle du déterminant sélectionnant un autre déterminant ou un nom, et celle de l'adjectif portant le trait lexical [$_{+postnominal}$], ce qui leur permet d'occuper des positions postnominales[2]. Quant aux pronoms clitiques, dans le cadre de notre analyse, ils constituent des pronoms, c'est-à-dire une sous-catégorie de N portant le trait lexical [$_{+CL}$], pouvant s'attacher aux noms, adjectifs, numéraux

[2] Notons que les adjectifs comme *όλος*, étant des adjectifs non-intersectifs, ne sont pas compatibles avec la polydéfinitude (cf. *ο κόσμος* (**ο*) *όλος*). Voir aussi note 4.

et quantifieurs[3]. La figure 3(a) présente la configuration du syntagme *το πρώτο καινούργιο αυτοκίνητό της* 'sa première voiture neuve' illustrant cette hypothèse ainsi que les possibilités du mouvement du clitique possessif et de son attachement à un adjectif (*καινούργιο της*) ou à un numéral (*πρώτο της*) en formant une chaîne clitique qui le lie à sa position initiale (*αυτοκίνητό της*). Plus précisément, Fips procède à la reconnaissance des mots du syntagme jusqu'à l'identification du nom *αυτοκίνητο* et à la projection de la tête clitique à sa droite. Dans le cas où le clitique s'est déplacé plus haut, l'analyseur crée une trace et en parallèle une coindexation du clitique avec sa position canonique. Quant à la forme nominale pleine du génitif possessif *της Ειρήνης* [4] 'd'Irène' dans le syntagme *το αυτοκίνητο της Ειρήνης* 'la voiture d'Irène', donné dans 3(b), elle occupe la position du complément de la tête nominale *το αυτοκίνητο* 'la voiture'. Son analyse est déterminée par une règle d'attachement du DP (+génitif) à droite du NP. La structure rendue par l'analyseur est la suivante : [$_{DP}$ *το* [$_{NP}$ *αυτοκίνητο* [$_{DP}$ *της* [$_{NP}$ *Ειρήνης*]]]]. La représentation arborescente de la structure en 3(b) illustre les deux possibilités de placement du DP possessif, soit à sa position canonique soit par déplacement dans le [Spec, DP] pour des raisons de focalisation.

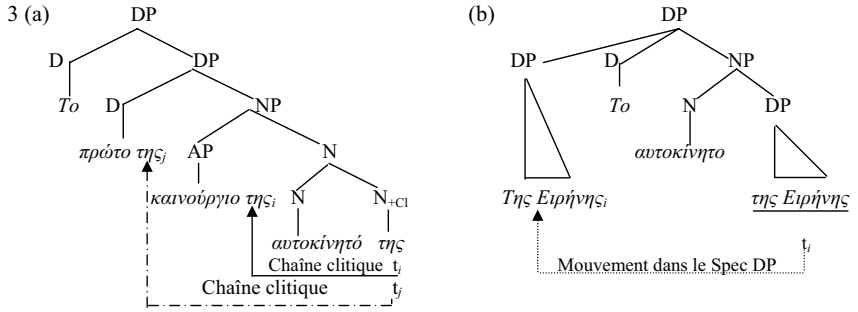


Figure 3(a-b) : Déplacement des clitiques et du DP possessif

Examinons maintenant le phénomène de la répétition du déterminant défini avec un modificateur adjectival prénominal (Det_{def}+Adj+Det_{def}+Nom) ou postnominal[5] (Det_{def} +Nom +Det_{def}+Adj). Vu que dans les deux cas les adjectifs ont une lecture prédicative, nous proposons un traitement relativement similaire, c'est-à-dire que Fips les analyse comme des structures FP de type *small clause*. Plus précisément, lors de la reconnaissance d'une séquence Det_{def} +Adj qui doit être impérativement précédée ou suivie d'une autre séquence Det_{def} +Nom, après la vérification de l'accord entre les deux séquences, nous proposons qu'une projection fonctionnelle FP se crée. Dans cette projection de prédication FP, l'article projetant un DP occupe la position du sujet et l'adjectif constitue le prédicat. La figure 4(a-b) illustre mieux cette configuration : dans la représentation du syntagme nominal *το αυτοκίνητο*

[3] Les éléments clitiques possessifs au génitif sont de la même nature (pronominale) que les éléments clitiques des phrases. A savoir, ils portent des traits de personne, ils ne font pas l'accord avec leur hôte et ils sont intransitifs.

[4] L'élément *της* est le Det_{def} et non le pronom clitique dont la forme est identique à l'article défini.

[5] La propagation de l'article défini n'est pas possible avec tous les adjectifs. En fait, les adjectifs dits non-intersectifs bloquent la polydéfinitude (Alexiadou & Wilder, 1998).

το μικρό 'la voiture la petite' la projection fonctionnelle s'attache à droite du NP (cf. 4(a)), contrairement à la représentation du syntagme *το μικρό το αυτοκίνητο* 'la petite la voiture' où le FP s'antépose en occupant la place du [Spec, DP] (cf. 4(b)).

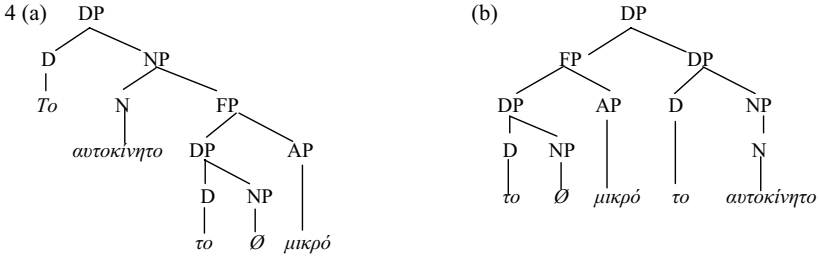


Figure 4(a-b) : Analyses proposées pour les structures polydéfinies

5 Evaluation

L'analyseur FipsGreek a été testé sur un ensemble de fichiers d'exemples des 1000 syntagmes nominaux manuellement extraits de textes en ligne. Ce petit corpus de test a été créé dans le but de nous permettre de contrôler chaque règle de grammaire relative à chaque structure syntaxique traitée. Notre lexique était déjà enrichi avec les nouveaux mots contenus dans ces fichiers. De ce fait, nous avons écarté le problème que les mots inconnus présentent dans l'analyse et nous avons obtenu le taux de 80% des syntagmes complètement analysées. Une analyse est dite complète si l'analyseur parvient à construire une structure arborescente des constituants couvrant l'intégralité de la phrase. Lorsqu'une analyse est incomplète l'analyseur retourne une séquence d'analyses partielles couvrant toute la phrase. Une évaluation manuelle des analyses complètes révèle que les analyses linguistiquement incorrectes représentent un très faible pourcentage de 4%. Les analyses incomplètes ou incorrectes sont dues à trois raisons principales. Premièrement, à l'ambigüité lexicale, l'analyseur n'ayant pas identifié la bonne catégorie des mots dans les syntagmes. Deuxièmement, au fait que dans les syntagmes non analysés, il existe des fautes d'orthographe, des mots étrangers, des abréviations et des sigles que l'analyseur n'a pas pu deviner. Troisièmement, au manque de la couverture de règles de la grammaire dans le système, fait qui est la cause de la majorité d'échecs. A savoir, il y a encore diverses structures que nous ne traitons pas, telles que les appositions, les constructions elliptiques, les figures du discours stylistiques, les dépendances à longue distance (Horrocks & Stavrou, 1987), etc.

6 Conclusion

Cet article présente une approche au traitement automatique des structures syntaxiques du syntagme nominal en grec moderne qui repose sur le postulat théorique de l'analyse DP. Cette analyse permet la sélection non seulement d'un NP mais aussi d'un DP par un autre DP et, par conséquent, des couches successives des DP que nous rencontrons dans le contexte du syntagme nominal en grec moderne. Notre proposition de modélisation utilise l'avantage que présente l'analyse hiérarchique du DP plutôt qu'une analyse plus linéaire avec une structure NP simple, pour la formalisation et l'implémentation des structures nominales du grec dans le cadre de l'analyseur multilingue Fips. Ce dernier est conçu avec des procédures et des modules généraux et des modules spécifiques au grec. Ainsi, FipsGreek, s'intégrant dans ce

système multilingue procède à une analyse de surface enrichie, parallèle à celles effectuées pour les autres langues, ce qui conduit à l'amélioration de la procédure du parsing dans le système de la traduction. Par ailleurs, les résultats obtenus montrent que notre modèle constitue un outil adéquat pour le traitement automatique des spécificités que présente le grec. La présente étude peut être étendue et couvrir l'analyse syntaxique de la phrase en donnant des perspectives à la recherche menée dans le futur.

Remerciements

Cette recherché a bénéficié du financement du Fonds national suisse de la recherche scientifique (projet n° 101412-103999). L'auteur tient à remercier Eric Wehrli, et Christopher Laenzlinger en particulier, pour leurs discussions et commentaires fructueux.

Références

ABNEY S. (1987). *The English Noun Phrase in its Sentential Aspect*. Cambridge, MA: MIT Press.

ALEXIADOU A., WILDER C. (1998). Adjectival modification and multiple determiners. In Alexiadou A. et al. (éd.), *Possessors, Predicates and Movement in the Determiner Phrase*, 303-332. Benjamins.

BALDZIS S.D., EUMERIDOU E., KOLALAS S. (2002). A Complete and Comprehensive System for Modern Greek Language Processing Proposed as a Modern Greek Language Call Method Developer. *Literary and Linguistic Computing* 17(4), 373-400.

ANDROUTSOPOULOU A. (1995). The distribution of definite determiners and the syntax of Greek DPs. In *Proceedings of CLS.*

DRAGGIOTIS A., GRIGORIADOU M., PHILOKYPROU G. (1998). The DINOUS parser. *Natural Language Engineering* 4(2), 145-173.

HOLTON D., MACKRIDGE P., PHILIPPAKI-WARBURTON I. (1997). *Greek: A Comprehensive Grammar of the Modern Language*. London: Routledge.

HORROCKS G., STAVROU M. (1987). Bounding theory and Greek syntax: evidence for wh- movement in NP. *Journal of Linguistics* 23, 79-108.

KERMANIDIS K., SGARBAS K., FAKOTAKIS N., KOKKINAKIS G. (2001). A PC-PATR-Based Syntactic Description of Modern Greek. *Literary and Linguistic Computing* 15(3), 291-311.

KOLLIAKOU D. (2004). Monadic definites and polydefinites: their form, meaning and use. *Journal of Linguistics* 40, 263-323.

LEONI DE LEÓN A., MICHOU A. (2006). Traitement des clitiques dans un environnement multilingue. Actes de la *13ème conference TALN 2006*, 541-550.

MARKOPOULOS G. (2001). *A Two-level Description of the Greek Noun Morphology with a Unification-based Word Grammar*. Ph.D. Dissertation: University of Athens.

STAVROU M. (1999). The position and serialization of APs in the DP: evidence from Greek. In Alexiadou A. et al. (éd.). *Studies in Greek Syntax,* 201-225. Dordrecht: Kluwer.

STAVROU M. (1996). Adjectives in Modern Greek. An instance of predication or an old issue revisited. *Journal of Linguistics* 32, 79-111.

TRIANTOPOULOU T. (1997). A Description of the Modern Greek Noun Phrase Using the "Affix Grammars over a Finite Lattice" Formalism. *Literary and Linguistic Computing* 12(2), 119-133.

RALLI A., GALIOTOU E. (1987). A morphological processor of Modern Greek. *Proceedings of the ACL European Chapter Meeting*. Denmark.

WEHRLI E. (2004). Un modèle multilingue d'analyse syntaxique. In Auchlin A. et al. (éd.), *Structures et discours : Mélanges offerts à Eddy Roulet*, 311-329. Canada: Nota bene.

Apprentissage symbolique de grammaires et traitement automatique des langues

Erwan MOREAU
LINA - FRE 2729, Université de Nantes
2 rue de la Houssinière, BP 92208, F-44322 Nantes cedex 3
Erwan.Moreau@univ-nantes.fr

Résumé. Le modèle de Gold formalise le processus d'apprentissage d'un langage. Nous présentons dans cet article les avantages et inconvénients de ce cadre théorique contraignant, dans la perspective d'applications en TAL. Nous décrivons brièvement les récentes avancées dans ce domaine, qui soulèvent selon nous certaines questions importantes.

Abstract. Gold's model formalizes the learning process of a language. In this paper we present the advantages and drawbacks of this restrictive theoretical framework, in the viewpoint of applications to NLP. We briefly describe recent advances in the domain which, in our opinion, raise some important questions.

Mots-clés : apprentissage symbolique, modèle de Gold, grammaires catégorielles.

Keywords: symbolic learning, Gold's model, categorial grammars.

1 Introduction

L'apprentissage symbolique automatique de grammaires pour les langues naturelles est un domaine relativement méconnu, assez peu étudié et très peu avancé sur le plan applicatif. Cet état de fait s'explique assez facilement : tout d'abord, il s'agit d'une tâche extrêmement complexe, aussi bien dans sa définition précise que dans sa mise en œuvre. Ce sont surtout les aspects théoriques qui en sont étudiés, et il semble jusqu'à présent très difficile d'y obtenir des résultats pratiques dignes d'intérêt (pour le langage naturel). D'un point de vue optimiste, la relative lenteur à passer du stade de l'étude théorique au stade des applications dans ce domaine s'explique par sa grande complexité. En ce sens, ce domaine serait simplement encore trop jeune scientifiquement, mais pourrait prendre de l'ampleur à l'avenir une fois que les bases en seront bien établies. Mais d'un point de vue pessimiste, la complexité excessive de la tâche peut être vue tout simplement comme un obstacle rédhibitoire à d'éventuelles applications.

Pourtant ce domaine est potentiellement riche en applications, si toutefois on admet l'hypothèse quelque peu idéaliste selon laquelle il est possible de construire un algorithme d'apprentissage « parfait ». Celui-ci serait donc capable de donner une grammaire précise d'un langage naturel, pourvu qu'on lui fournisse un nombre suffisant de phrases appartenant à celui-ci. Sous cette hypothèse, la première application (et la plus évidente) est l'analyse syntaxique, elle-même utilisée sous différentes formes dans de nombreux outils de traitement des langues. On pourrait

alors envisager de construire assez facilement des analyseurs, y compris pour des langues pour lesquelles peu d'études linguistiques existent. On peut également penser à coupler l'analyse et l'apprentissage, de façon à mieux prendre en compte la catégorie syntaxique des mots inconnus de l'analyseur. D'autres applications liées à l'analyse, telles que la correction orthographique et syntaxique, sont également susceptibles de bénéficier des apports de l'apprentissage automatique. Si l'on dispose aussi des moyens permettant de gérer l'aspect sémantique des langues, l'autre grande application de l'apprentissage est la génération (passage du sens d'un énoncé à sa réalisation dans une langue précise). Celle-ci est elle-même proche du problème de la traduction automatique, qui est bien sûr une application d'une très grande utilité.

L'*inférence grammaticale* désigne la problématique qui consiste à apprendre des langages à partir de données. Tout cadre formel pour ce problème doit donc avant tout définir les termes *apprentissage*, *langages* et *données*, c'est-à-dire répondre aux questions suivantes : nature des données dont on dispose ? simples séquences de mots (chaînes), arbres, termes, graphes ou tout autre type de structures, mais aussi quantité, qualité, complétude des données. Type de langages considéré, et représentation des langages ? restrictions éventuelles, niveau d'abstraction (e.g. sans contrainte particulière sur la relation entre langages et grammaires, ou au contraire formalisme grammatical précis). Nature du processus d'inférence ? Fini ou non, Solution unique ou multiple, processus automatique ou semi-automatique, résultat précis ou approximation, limites éventuelles sur le temps ou le nombre d'essais.

Le modèle d'identification à la limite, aussi appelé du nom de son auteur modèle de Gold, est l'une des principales représentations formelles du processus d'apprentissage. La première définition en est donnée dans (Gold, 1967). L'auteur lui-même est d'abord pessimiste quant à l'intérêt de ce modèle, à cause de l'apparente impossibilité d'y obtenir des résultats positifs pour des classes de langages « intéressantes ». Plus tard, les résultats positifs obtenus par Angluin dans ce modèle montreront sa pertinence (Angluin, 1980). Le modèle sera ensuite étudié plus en détail : ainsi, plusieurs autres résultats encourageants viendront soutenir l'idée que l'identification à la limite constitue bien un cadre théorique adapté à la représentation du processus d'apprentissage, en particulier celui des langages, voire des langues naturelles.

La question de la pertinence du modèle de Gold par rapport à l'acquisition humaine du langage est plutôt bien étudiée au niveau linguistique et cognitif (Johnson, 2004), mais cette même question est assez peu discutée dans la perspective de l'apprentissage symbolique automatique. C'est pourquoi nous proposons ici une relecture des principaux résultats liés à ce modèle, vu sous l'angle du traitement automatique des langues. Dans cet article nous essaierons donc d'expliquer de façon concise et claire les bases, les outils et les enjeux du modèle de Gold par rapport au TAL. Nous proposons ensuite un point de vue particulier sur les intérêts et limites des résultats obtenus jusqu'ici dans ce cadre, en tentant de donner un peu de recul à cette modeste étude. L'objectif de cet article est donc aussi de soumettre à la discussion quelques questions relatives au domaine, qui nous semblent pertinentes compte tenu de ses récentes évolutions.

2 Identification à la limite

Le principe de l'identification à la limite est la convergence : à partir d'une séquence infinie d'éléments qui caractérisent le langage à deviner, l'apprenant émet des hypothèses. Ces hypothèses prennent la forme d'une grammaire, censée correspondre au langage observé jusqu'alors par l'apprenant. Comme l'énumération est infinie, l'apprenant répond lui aussi sous forme d'une

séquence infinie de grammaires hypothèses. Finalement, l'apprentissage est réussi si, à partir d'un certain point, l'apprenant émet toujours la même hypothèse (convergence), et que celle-ci correspond bien au langage attendu. Le fait que l'apprenant ignorera toujours s'il a atteint ou non la solution est un aspect important de ce formalisme. Gold en donne la justification (d'ordre linguistique) suivante : *« une personne ne sait jamais si elle parle correctement un langage. »*.

Une *classe de langages* est un ensemble de langages[1] fixé, parfois aussi appelé *famille de langages*. Généralement il s'agit d'un ensemble de langages partageant une propriété particulière. L'*apprenabilité*[2] d'une classe de langages désigne son aptitude à être *apprise* selon la définition 2.1 ci-dessous.

Un *système de grammaires* est spécifié par un triplet $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$, dans lequel l'univers $\mathcal{U}$ est un ensemble d'objets, $\mathcal{G}$ un ensemble de grammaires et $\mathcal{M}$ une fonction qui associe à chaque grammaire de $\mathcal{G}$ un sous-ensemble de $\mathcal{U}$. Dans un système de grammaires $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$, une *fonction d'apprentissage* est une fonction partielle ϕ qui associe à des séquences finies non-vides d'objets de $\mathcal{U}$ des grammaire de $\mathcal{G}$.

Définition 2.1 (Identification à la limite) *Soit $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$ un système de grammaires, ϕ une fonction d'apprentissage et $L \subseteq \mathcal{U}$ un langage. Soit $\langle a_0, a_1, a_2, \ldots \rangle$ une séquence infinie d'objets de $\mathcal{U}$, telle que $a \in \{ a_i \mid i \in \mathbb{N} \}$ si et seulement si $a \in L$.*

ϕ converge *vers G s'il existe $n \in \mathbb{N}$ tel que pour tout $i \geq n$ $G_i = \phi(\langle a_1, a_2, \ldots, a_i \rangle)$ est définie et $G_i = G$.*

ϕ apprend *un langage L si, pour toute énumération de L, ϕ converge vers une grammaire G telle que $\mathcal{M}(G) = L$.*

Une classe de langages $\mathcal{L} \subseteq \mathcal{P}(\mathcal{U})$ est dite apprenable *s'il existe une fonction d'apprentissage ϕ telle que ϕ apprend L pour tout langage $L \in \mathcal{L}$.*

On voit dans cette définition que la séquence d'exemples a quelques caractéristiques notables :

- Elle ne contient que des exemples *positifs*, c'est-à-dire des éléments du langage. La fonction d'apprentissage n'a donc aucune information extérieure au langage, ce qui constitue la principale difficulté de cette forme d'apprentissage.
- La séquence d'exemples est supposée ne comporter aucune erreur[3].
- La séquence d'exemples est une *énumération* du langage : tous les objets du langage doivent obligatoirement y apparaître.
- Les exemples peuvent apparaître dans un ordre quelconque dans la séquence, et éventuellement plusieurs fois (ce qui permet notamment d'énumérer indéfiniment un langage fini).

Dans ce modèle, la convergence d'une fonction d'apprentissage n'a d'intérêt que si elle s'applique à un ensemble de langages, et non à un seul langage. Intuitivement, plus la classe de langages est grande, plus il est difficile de reconnaître un langage précis dans cette classe.

[1] Dans la littérature, le terme *langage* est fréquemment défini comme un ensemble de phrases, chaque phrase étant une séquence finie de mots. Mais dans la mesure où on peut envisager différents niveaux de représentation de la phrase, nous définissons un *langage* comme un ensemble d'*objets* (ce terme abstrait laissant volontairement la possibilité d'utiliser différents types d'éléments : arbres, structures, etc.).

[2] On trouve aussi dans la littérature différents termes désignant le caractère apprenable d'une classe de langages : *inférable*, *identifiable [à la limite]* ou *acquiérable* (le terme *acquisition* fait cependant plus souvent référence à l'apprentissage humain du langage).

[3] Ce qui limite les applications potentielles : ce modèle est par définition inadapté aux données bruitées.

Cette définition de l'identification à la limite a d'importantes conséquences immédiates, démontrées par Gold dans (Gold, 1967). La première est un résultat positif pour les langages finis : La classe des langages de cardinalité finie est apprenable. En effet, intuitivement il suffit dans ce cas que l'apprenant ajoute un par un les exemples présentés à la grammaire hypothèse : la fonction d'apprentissage converge dès que le langage a été énuméré en totalité. En revanche la seconde conséquence de la définition du modèle est un résultat négatif : toute classe de langages contenant tous les langages finis et au moins un langage infini n'est pas apprenable. Nous illustrons ce résultat à l'aide de l'exemple ci-dessous :

Exemple 2.1 (Langages réguliers) *Pour tout $n \geq 1$ on définit $L_n = \{ x^i \mid i \leq n \}$ comme le langage des chaînes de x de longueur inférieure ou égale à n. Soit $L_\infty = x^*$ le langage de toutes les chaînes de x.*

Supposons que la séquence d'exemples commence par $\langle x, xx, xxx, \ldots \rangle$:
- *Si l'apprenant est prudent, il ne propose jamais comme hypothèse un langage qui va au delà des exemples proposés : il propose donc L_k, avec k la longueur maximale parmi les exemples vus. Cet algorithme ne peut jamais trouver L^∞.*
- *Si à l'inverse l'algorithme « généralise », alors à partir d'un certain point il propose L^∞. Mais c'est une erreur s'il s'avère que la séquence ne dépasse pas une certaine longueur de phrase.*

Une erreur est donc possible dans les deux cas, et rien ne permet de faire le bon choix : si la classe de langage contient tous les L^n et L^∞, celle-ci n'est pas apprenable.

Les langages $\{ L_n \mid n \in \mathbb{N} \}$ et L_∞ définis dans l'exemple 2.1 étant tous réguliers, toutes les classes de la hiérarchie de Chomsky les contiennent, et ne sont par conséquent pas apprenables. Le fait que même la classe la plus simple de la hiérarchie de Chomsky, celle des langages réguliers, ne soit pas apprenable dans le modèle de Gold a longtemps constitué un obstacle majeur au développement du modèle, considéré comme trop contraignant. Gold lui-même notait : *« Cependant, les résultats présentés dans la dernière section montrent que seule la classe de langages la plus triviale* [4] *considérée*[5] *est apprenable à partir d'exemples positifs[..]. »*

L'exemple 2.1 illustre la principale difficulté de l'apprentissage à partir d'exemples positifs, à savoir la *surgénéralisation*. La surgénéralisation est l'erreur qui consiste à trop généraliser (extrapoler) à partir des données fournies, ce qui signifie inférer un langage qui est un sur-ensemble strict du langage cible. Par exemple, on peut supposer que l'ensemble $\{11, 23, 5, 17, 7\}$ est le début d'une énumération de l'ensemble des nombres impairs. Mais s'il s'agit en fait de l'ensemble des nombres premiers supérieurs à 2, alors il y a surgénéralisation : l'ensemble des nombres représenté est un sous-ensemble (strict) de l'ensemble proposé. Comme on ne dispose que d'exemples positifs, il n'y aura jamais de contre-exemple dans la séquence permettant de corriger l'erreur. Bien entendu, la généralisation est indispensable dans le processus d'apprentissage, puisqu'une méthode d'apprentissage « trop prudente » qui ne généraliserait jamais ne ferait pas de véritable *apprentissage* (au sens d'une découverte de quelque chose de nouveau) : il s'agirait simplement d'une sorte de compilation des exemples proposés. Surtout, il est évident qu'une telle méthode serait incapable d'identifier un langage infini.

Sauf cas particuliers, la généralisation doit donc bien être utilisée au cours de l'apprentissage. La question qui se pose d'un point de vue algorithmique est : quand faut-il généraliser ? (ou

[4] On peut considérer de manière informelle qu'une classe de langages est *non triviale* (pour l'apprentissage) si elle comporte au moins un nombre infinis de langages, dont certains sont infinis.

[5] Il s'agit de la classe des langages de cardinalité finie.

quand faut-il ne pas généraliser, selon ce qu'on considère comme étant l'action par défaut). Mais avant de se poser cette question, il faut s'assurer qu'il est *possible de savoir quand généraliser*, car lorsqu'on ne dispose pas d'exemples négatifs on n'a aucun indice sur la position de la frontière du langage à deviner. C'est précisément ce point qui pose problème au départ avec l'identification à la limite à partir d'exemples positifs : même les classes de langages qu'on croyait simples (les langages réguliers, voir exemple 2.1) ne sont pas apprenables, parce qu'on ne peut pas savoir quand [ne pas] généraliser.

3 Techniques théoriques d'apprentissage de grammaires

3.1 Ensembles révélateurs

Au début des années 1980, Angluin apporte au modèle de Gold ses premiers résultats positifs : dans (Angluin, 1980), elle propose de *« considérer le cas particulier d'inférence à partir d'exemples positifs qui évite la surgénéralisation [et donne] des conditions suffisantes pour cela. »* Le critère qu'elle propose a donné lieu ensuite à de nombreuses utilisations ou extensions, démontrant finalement la richesse du modèle de Gold. Sommairement, un ensemble révélateur *(telltale set)* est une sorte de « signature » d'un langage qui le distingue de tous les autres langages de la classe dont il est un sur-ensemble strict. Ainsi, lorsque cette signature apparaît dans la séquence d'exemples, on peut proposer ce langage sans risque de surgénéralisation. Formellement, soit $\mathcal{L}$ une classe de langages : un ensemble fini d'objets D est un *ensemble révélateur* du langage $L \in \mathcal{L}$ si $D \subseteq L$ et $L' \subset L \Rightarrow D \nsubseteq L'$ pour tout langage $L' \in \mathcal{L}$.

Théorème 3.1 (Angluin) *Soit $\mathcal{L} \subseteq \mathcal{P}(\mathcal{U})$ une famille indexée de langages récursifs*[6] *dans le système de grammaires $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$: $\mathcal{L} = \{ \mathcal{M}(G) \mid G \in \{G_0, G_1, G_2, \ldots\} \}$.*

Il existe une fonction ϕ qui apprend $\mathcal{L}$ si et seulement s'il existe un algorithme calculable qui, pour tout indice I tel que $L_I = \mathcal{M}(G_I) \in \mathcal{L}$ énumère un ensemble révélateur de L_I.

Supposons qu'il existe un algorithme $EnumRevel(I, n)$ qui énumère récursivement les n premiers éléments d'un ensemble révélateur D_I de L_I.

$\phi(\langle a_0, \ldots, a_n \rangle)$
  $i \leftarrow 0$
  $E \leftarrow EnumRevel(i, n)$
  Tant que ($i \leq n$ **et non**($\{a_0, \ldots, a_n\} \subseteq \mathcal{M}(G_i)$ **et** $E \subseteq \{a_0, \ldots, a_n\}$)) **faire**
    $i \leftarrow i + 1$
    $E \leftarrow EnumRevel(i, n)$
  Fin Tant Que
  Renvoyer G_i

L'algorithme ci-dessus apprend la classe $\mathcal{L}$ de la façon suivante : un ensemble révélateur étant nécessairement fini, pour tout i il existe une étape n à partir de laquelle l'ensemble révélateur E de L_i est énuméré en totalité. Dans ce cas, la boucle s'arrêtera sur la première grammaire G_i telle que les exemples fournis appartiennent au langage de la grammaire d'une part, et dont

[6] Classe de langage pour laquelle le problème de l'appartenance ($x \in L$) est décidable.

l'ensemble révélateur est entièrement inclus dans les exemples d'autre part. Ainsi il est impossible que le langage cible soit un sous-ensemble strict de $\mathcal{M}(G_i)$ (surgénéralisation). Si aucun contre-exemple $a_j \notin \mathcal{M}(G_i)$ n'apparaît par la suite, l'algorithme s'arrêtera toujours sur G_i.

L'apprentissage par énumération, illustré ci-dessus, désigne une méthode générale qui consiste à faire une recherche systématique dans l'ensemble des grammaires possibles, jusqu'à en trouver une qui vérifie une propriété particulière, et la renvoyer comme hypothèse. L'énumération n'est pas une méthode d'apprentissage universelle, parce qu'il existe des classes de langages apprenables qui ne sont pas apprenables par énumération (Costa Florêncio, 2003). Il va de soi que ce type d'algorithme est totalement inutilisable en pratique : même si on peut améliorer sensiblement la méthode (notamment en évitant de faire l'énumération complète après chaque exemple), le seul fait d'avoir à parcourir de façon exhaustive l'ensemble des grammaires potentielles est rédhibitoire. En effet, imaginons une représentation textuelle simple des grammaires (de type grammaires syntagmatiques), de façon à les énumérer selon l'ordre de taille puis lexicographique : en première approximation il existe de l'ordre de n^m grammaires différentes de taille m, avec n le nombre total de symboles (comprenant entre autres tous les mots du vocabulaire). Ce type de fonctionnement est évidemment radicalement inadapté à l'apprentissage de langues naturelles[7]. Il faut donc souligner cette caractéristique essentielle du modèle de Gold : il s'agit avant tout d'un modèle *théorique*, qui ne garantit que la *décidabilité* du problème de l'apprentissage. Autrement dit, le fait qu'une classe de langages soit apprenable n'implique en aucun cas la faisabilité pratique (en temps raisonnable) du processus d'apprentissage.

3.2 Élasticité finie

L'élasticité [in]finie est une propriété définie par (Motoki *et al.*, 1991) de la façon suivante : Soit $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$ un système de grammaires. Une classe $\mathcal{L} \subseteq \mathcal{P}(\mathcal{U})$ a l'élasticité infinie s'il existe une séquence infinie $\langle a_0, a_1, a_2, \ldots \rangle$ d'objets dans $\mathcal{U}$ et une séquence infinie $\langle L_1, L_2, \ldots \rangle$ de langages dans $\mathcal{L}$ tels que $a_i \notin L_i$ et $\{a_0, \ldots, a_{i-1}\} \subseteq L_i$ pour tout $i > 0$. Une classe de langages $\mathcal{L}$ a la propriété d'*élasticité finie* si elle n'a pas l'élasticité infinie. L'élasticité finie est donc une condition suffisante pour l'apprenabilité. De fait, il s'agit d'une propriété très utile pour démontrer l'apprenabilité de nouvelles classes de langages, car cette condition est souvent plus simple à vérifier que l'existence globale d'un algorithme convergent. L'élasticité finie est notamment utilisée par Shinohara pour définir une nouvelle condition suffisante à l'aide du concept de *densité finie bornée*, qui lui permet de démontrer l'apprenabilité de la classe des grammaires syntagmatiques contextuelles d'au plus k règles (pour tout $k \geq 0$) (Shinohara, 1991). De plus, Kanazawa a démontré une propriété très pratique, qui permet de montrer l'élasticité finie (donc aussi l'apprenabilité) d'une classe de langages complexe à partir du cas d'une classe plus simple possédant la propriété (voir ci-dessous). D'un point de vue algorithmique, on notera que tous les résultats d'apprenabilité obtenus à l'aide de l'élasticité finie reposent finalement sur l'algorithme d'apprentissage par énumération des ensembles révélateurs (présenté plus haut).

Théorème 3.2 (Kanazawa) *Soient $\mathcal{U}$ et $\mathcal{U}'$ deux ensembles d'objets, et $\mathcal{L}$ une classe de langages définie sur $\mathcal{U}$ qui a l'élasticité finie. S'il existe une relation $R \subseteq \mathcal{U}' \times \mathcal{U}$ finiment valuée, alors la classe de langages $\mathcal{L}' = \{\ R^{-1}[L] \mid L \in \mathcal{L}\ \}$ a aussi l'élasticité finie*[8].

[7] Notons que l'acquisition humaine du langage n'a certainement rien à voir non plus avec cette méthode.

[8] Une relation binaire R sur $A \times B$ est *finiment valuée* si et seulement si pour tout $a \in A$ il n'existe qu'un nombre fini de $b \in B$ tels que $a\ R\ b$. Si L est un langage sur $\mathcal{U}$ et R une relation sur $\mathcal{U}' \times \mathcal{U}$, l'*image inverse* de L par rapport à R est le langage $R^{-1}[L] = \{\ a \in \mathcal{U}' \mid \exists b \in L \text{ tel que } a\ R\ b\ \}$.

4 Apprentissage de grammaires catégorielles

En 1998, Kanazawa propose plusieurs résultats importants concernant l'apprenabilité des grammaires AB dans le modèle de Gold (Kanazawa, 1998). Les grammaires AB, la forme la plus simple de grammaires catégorielles, sont (totalement) lexicalisées : à chaque mot sont associés un ou plusieurs types syntaxiques dans le lexique (règles *lexicales*), et deux règles *universelles* définissent la façon dont ces types peuvent se combiner entre eux dans les dérivations[9].

Les apports de Kanazawa sont multiples : il montre de nouveaux résultats et développe de nouvelles techniques de preuve. Surtout, ses résultats sont les premiers pour le modèle de Gold à traiter d'un formalisme grammatical présentant certaines prédispositions à la représentation des langues naturelles, à savoir les grammaires catégorielles. Plus précisément, les grammaires AB sont assez pauvres sur le plan de la représentation linguistique. Mais la famille des grammaires catégorielles contient d'autres formalismes beaucoup plus puissants pour représenter des langues naturelles, c'est pourquoi le premier résultat prometteur de Kanazawa a donné lieu à d'autres travaux visant à étendre l'apprenabilité à des formes plus riches de grammaires.

4.1 Apprenabilité des grammaires AB

Parmi ses résultats, il faut distinguer deux aspects très différents du point de vue applicatif :

Il y a tout d'abord un aspect algorithmique, basé sur l'algorithme d'apprentissage RG proposé dans (Buszkowski & Penn, 1989). Cet algorithme apprend efficacement des grammaires AB rigides[10] à partir de FA-structures. Ces structures sont une forme « d'arbre de dérivation appauvri » des phrases, c'est-à-dire qu'elles ne contiennent pas toutes les informations d'un arbre de dérivation (sans quoi il n'y aurait aucun apprentissage, puisque les types seraient déjà donnés), mais tout de même beaucoup plus d'information que de simples chaînes : parenthésage des constituants, ainsi qu'une forme particulière d'orientation des dépendances entre constituants. Il est donc plus facile d'apprendre lorsqu'on dispose en plus de cette information structurée.

Le second aspect concerne l'apprenabilité d'une classe de langages plus étendue. Kanazawa ne montre pas seulement l'apprenabilité de la classe des langages de FA-structures de grammaires AB rigides, il démontre aussi que cette classe a l'élasticité finie. Or grâce au théorème 3.2, il prouve que cette propriété est également vérifiée par la classe des langages de chaînes des grammaires AB k-valuées[11] (pour tout $k \geq 0$), donc cette classe est elle aussi apprenable. Ce résultat est beaucoup plus intéressant pour deux raisons : d'une part la contrainte de rigidité est levée, ce qui permet d'envisager de représenter un langage naturel avec ces grammaires[12]. D'autre part il n'est plus nécessaire de disposer des FA-structures avec les exemples de phrases, ce qui est un avantage important puisque celles-ci constituent une information spécifique au formalisme, en pratique très difficile à obtenir en quantité. En revanche, on ne dispose pas dans ce cas d'algorithme d'apprentissage efficace[13].

[9] Voir par exemple (Moreau, 2006) pour une définition complète.

[10] Une grammaire est rigide si à chaque mot du vocabulaire n'est associé qu'un seul type syntaxique.

[11] Une grammaire est k-valuée si à chaque mot du vocabulaire n'est associé qu'au plus k types différents.

[12] La rigidité empêche en effet toute forme d'homonymie. Mais surtout elle ne permet pas de représenter de manière satisfaisante la plupart des mots grammaticaux, car leur usage syntaxique prend souvent des formes variés.

[13] Au contraire, Costa-Florêncio démontre qu'il s'agit d'un problème NP-dur (Costa Florêncio, 2003).

4.2 Extensions à d'autres formalismes

Les bons résultats obtenus par Kanazawa avec les grammaires AB posent la question de savoir si les grammaires catégorielles ont certaines propriétés qui feraient d'elles de bonnes candidates à l'apprentissage dans le modèle de Gold. Cette question du formalisme grammatical est importante pour d'éventuelles applications aux langues naturelles, puisque celles-ci nécessitent une représentation à la fois linguistiquement fiable et aussi utilisable le plus facilement possible. En ce qui concerne l'apprenabilité efficace à partir de structures (de type FA-structures, mais la forme peut varier selon les formalismes), plusieurs résultats viendront montrer ensuite que ce type d'apprentissage peut être étendu à d'autre formalismes sans grande difficulté. Kanazawa donne lui-même l'exemple des grammaires combinatoires générales (GCG). Des résultats équivalents sont obtenus avec différents formalismes, notamment les grammaires de Lambek et les grammaires minimalistes (Bonato & Retoré, 2001), mais toujours au prix d'une contrainte similaire à la rigidité (limitations sur le nombre ou la forme des règles lexicales associées à un mot), et toujours avec l'aide d'informations structurelles assez précises.

Mais le passage du cas « grammaires rigides et avec structures » au cas « grammaires k-valuées ou sans structure », qui constitue le point fort des résultats de Kanazawa, s'avère nettement plus difficile lorsqu'on s'éloigne du cas des grammaires AB. On aurait pu supposer que les propriétés logiques des grammaires AB jouaient un rôle pour l'apprenabilité, mais cette hypothèse est invalidée par les résultats négatifs des grammaires de Lambek (Foret & Le Nir, 2002). Une autre hypothèse de travail a consisté à considérer les grammaires AB comme un système de grammaires (lexicalisées) spécifié par un ensemble particulier de règles universelles (de réécriture par substitution). On peut alors étudier ce qui les distingue des autres systèmes dans le cadre plus large des GCG proposé par Kanazawa (Moreau, 2006) : on cherche ainsi des conditions, portant sur les règles universelles, qui sont suffisantes pour l'apprenabilité des classes de langages correspondantes (on espère trouver de cette manière des ensembles de règles plus fines qui permettent l'apprenabilité). En se basant sur la méthode employée par Kanazawa, nous avons ainsi montré que certaines classes de GCG ont l'élasticité finie (donc sont apprenables) : les *grammaires à arguments bornés* k-valuées, qui représentent des classes de langages assez vastes, mais souffrent d'une limitation « technique » (sur la taille des arguments) difficile à justifier au niveau linguistique. Les *grammaires par consommation stricte d'arguments* k-valuées sont en revanche apprenables sans limitation, mais sont définies par un critère tellement strict qu'on ne s'éloigne pas beaucoup du cas des grammaires AB. De plus, il ne s'agit pas que d'une limite « temporaire » (c'est-à-dire susceptible d'être repoussée à l'avenir) car les *grammaires par consommation d'arguments* (rigides), qui en sont un sur-ensemble très peu élargi, n'ont pas l'élasticité finie : cela signifie qu'on atteint ici, entre ces deux cas relativement proches, les frontières de l'apprenabilité des GCG (du moins selon la méthode de Kanazawa).

4.3 Applications à l'apprentissage symbolique du langage naturel ?

Compte tenu des contraintes du modèle et des résultats présentés ci-dessus, il est compréhensible que les applications de ce type d'apprentissage au langage naturel demeurent très modestes. De fait, le premier problème à résoudre est cette équation apparemment insoluble : soit on cherche à apprendre à l'aide d'informations structurées, mais le type d'information requis n'existe pas en quantité suffisante a priori ; ou bien on tente d'apprendre à partir de simples phrases, mais alors on ne dispose que d'algorithmes de complexité exponentielle, incapables de réaliser le processus en temps raisonnable.

Différentes méthodes ont été envisagées, qui font toutes appel à des ressources structurées, de façon plus ou moins directe. Quelques unes utilisent des corpus de structures spécifiques, obtenus manuellement ou par conversion plus ou moins automatique de ressources arborées existantes (Dudau-Sofronie, 2004). Nous avions également proposé une approche intermédiaire, à partir de chaînes mais avec l'apport d'un sous-ensemble de la grammaire cible (Moreau, 2006), en utilisant un lexique existant sous forme de grammaires de liens. Le fait qu'il soit nécessaire de faire appel à des ressources externes, souvent exprimées dans un formalisme grammatical particulier, pose un problème théorique de fond du point de vue du modèle de Gold : où s'arrête la notion d'inférence grammaticale, c'est-à-dire d'apprentissage symbolique de la syntaxe, et où commence la « simple » extraction d'informations syntaxiques ? En effet, l'usage de ressources externes facilite bien sûr l'apprentissage, mais introduit aussi un biais dans le processus : à partir d'un certain niveau d'informations syntaxiques fournies, il ne s'agit plus d'apprentissage mais de reconstitution de la grammaire qui a servi à produire les exemples, qu'elle soit formellement établie ou sous-jacente. On risque alors de ne faire que reproduire des schémas syntaxiques préétablis, la grammaire résultante n'aurait donc pas beaucoup d'intérêt : dans ce cas elle peut être construite directement de façon semi-automatique, à partir des règles qui ont défini la création des données. Un autre travers plus subtil peut également apparaître : le simple étiquetage syntaxique par des catégories prédéfinies (nom, verbe, adjectif, etc.) est une forme appauvrie d'apprentissage de la syntaxe, car ce cadre empêche de tenir compte d'éventuelles variations par rapport aux catégories de départ. Dans ce cas, il n'est pas certain que le modèle de Gold ait quelque chose de plus à apporter au problème que les techniques existantes en TALN.

5 Conclusion

L'acquisition automatique de grammaires ne se limite pas à l'apprentissage (au sens de Gold). Par exemple, pour certaines formes de grammaires catégorielles, les travaux d'Hockenmaier (Hockenmaier, 2003) ou de Moot (Moortgat & Moot, 2001) montrent qu'il est possible d'obtenir une grammaire à large couverture d'un langage naturel, à partir de corpus structurés. Mais leur approche est à notre sens plus proche de l'*extraction* automatique que de l'inférence grammaticale, car dans les deux cas des techniques ad hoc de conversion des données sont utilisées.

L'utilisation d'un modèle contraignant comme le modèle de Gold constitue une garantie de « précision » de la grammaire obtenue, parce qu'il donne une direction générale au processus de l'apprentissage : l'existence d'un objectif (qu'on peut considérer comme idéal) définissant *ce que doit être* la grammaire apprise diffère de la simple extraction d'information syntaxique, dans laquelle on obtient toujours un résultat (quelles que soient les données), et ce résultat n'est justifié qu'a posteriori (parfois selon une évaluation spécifique, souvent simplement par son utilité). Typiquement, le problème de la surgénéralisation est difficile voire impossible à détecter dans le cas de l'extraction, tandis que le modèle de Gold impose d'en tenir compte *a priori* dans l'algorithme d'apprentissage (sans quoi la convergence ne serait pas vérifiée).

Il est vrai que le modèle de Gold est avant tout un modèle théorique, et le critère de convergence sur lequel il repose ne semble pas vraiment approprié pour des applications de traitement automatique. Dans (Angluin & Smith, 1983), Angluin concluait son état de l'art sur l'inférence inductive par la remarque suivante : *« Le problème ouvert le plus important n'est sans doute pas une quelconque question technique spécifique, mais le fossé entre les résultats abstraits et concrets. »* Force est de constater que, malgré quelques progrès indéniables sur le plan théorique, les tentatives d'applications concrètes de cette forme d'apprentissage restent encore peu

concluantes, parce qu'on ne parvient pas à (on ne peut pas ?) apprendre sur des données réelles sans relâcher tout ou partie des contraintes du modèle. Cela ne signifie pas nécessairement que l'on perde ainsi tout l'intérêt du modèle, mais dans ces conditions il nous semble judicieux de redéfinir l'objectif de la tâche d'apprentissage : inférence grammaticale, extraction, ou approche mixte ? Étant donné les difficultés rencontrées lorsqu'on s'en tient strictement au modèle, cette dernière possibilité semble la plus réaliste.

Toutefois, peut-être que l'algorithme d'apprentissage idéal n'est tout simplement pas encore découvert : dans ce cas, *« les générations futures riront bien de notre ignorance actuelle. »* (Angluin & Smith, 1983).

Références

ANGLUIN D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, **48**, 117–135.

ANGLUIN D. & SMITH C. H. (1983). Inductive inference : Theory and methods. *ACM Computing Surveys*, **15**(3), 237–269.

BONATO R. & RETORÉ C. (2001). Learning rigid lambek grammars and minimalist grammars from structured sentences. In *Proc. of 3d Workshop on Learning Language in Logic*, p. 23–34.

BUSZKOWSKI W. & PENN G. (1989). *Categorial grammars determined from linguistic data by unification*. Rapport interne TR-89-05, Dpt of Computer Science, University of Chicago.

COSTA FLORÊNCIO C. (2003). *Learning categorial grammars*. PhD thesis, Utrecht University.

DUDAU-SOFRONIE D. (2004). *Apprentissage de grammaires catégorielles pour simuler l'acquisition du langage naturel à l'aide d'informations sémantiques*. PhD thesis, Univ. Lille 1.

FORET A. & LE NIR Y. (2002). Lambek rigid grammars are not learnable from strings. In *COLING'2002, 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

GOLD E. (1967). Language identification in the limit. *Information and control*, **10**(5), 447–474.

HOCKENMAIER J. (2003). *Data and models for statistical parsing with Combinatory Categorial Grammar*. PhD thesis, School of Informatics, The University of Edinburgh.

JOHNSON K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, **71**, 571–592.

KANAZAWA M. (1998). *Learnable classes of categorial grammars*. Cambridge University Press.

MOORTGAT M. & MOOT R. (2001). CGN to Grail : Extracting a type-logical lexicon from the CGN annotation. In *Proceedings of CLIN 2000* : W. Daelemans.

MOREAU E. (2006). *Acquisition de grammaires lexicalisées pour les langues naturelles*. PhD thesis, Université de Nantes.

MOTOKI T., SHINOHARA T. & WRIGHT K. (1991). The correct definition of finite elasticity : corrigendum to Identification of unions. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, p. 375, San Mateo, CA : Morgan Kaufmann.

SHINOHARA T. (1991). Inductive inference of monotonic formal systems from positive data. *New Generation Computing*, **8**(4), 371–384.

Méthodes d'alignement des propositions : un défi aux traductions croisées

Yayoi NAKAMURA-DELLOYE
Université Paris 7 – LATTICE
1 rue Maurice Arnoux 92120 Montrouge
http://www.lattice.cnrs.fr
yayoi@free.fr

Résumé. Le présent article décrit deux méthodes d'alignement des propositions : l'une basée sur les méthodes d'appariement des graphes et une autre inspirée de la classification ascendante hiérarchique (CAH). Les deux méthodes sont caractérisées par leur capacité d'alignement des traductions croisées, ce qui était impossible pour beaucoup de méthodes classiques d'alignement des phrases. Contrairement aux résultats obtenus avec l'approche spectrale qui nous paraissent non satisfaisants, l'alignement basé sur la méthode de classification ascendante hiérarchique est prometteur dans la mesure où cette technique supporte bien les traductions croisées.

Abstract. The present paper describes two methods for clauses alignment. The first one uses a graph matching approach, while the second one relies on agglomerative hirerarchical clustering (AHC). Both methods are characterized by the fact they can align cross translations, which was impossible for previous classic sentence alignment methods. Though the results given by the spectral method are unsatisfactory, the method based on AHC is very promising. It handles correctly cross translations.

Mots-clés : alignement des corpus parallèles, appariement de graphes, classification ascendante hiérarchique, proposition syntaxique, mémoire de traduction, linguistique contrastive.

Keywords: parallel corpora alignment, graph matching, agglomerative hierarchical clustering, syntactic clause, translation memory, contrastive linguistics.

1 Introduction

L'alignement des propositions consiste en la mise en correspondance des propositions syntaxiques avec leurs traductions dans des textes parallèles. Les corpus parallèles alignés au niveau des propositions pourraient être des ressources profitables dans beaucoup de domaines tels que la traduction automatique ou la traduction assistée par ordinateur (TAO), ainsi que pour les recherches en linguistique contrastive. En dépit de cet intérêt notable, peu de travaux sur l'alignement des propositions ont été réalisés et les quelques méthodes proposées sont semblables à celles pour l'appariement des phrases. Or l'alignement des propositions entre deux langues relativement différentes sur tous les plans, telles que la paire français-japonais, n'est pas réalisable par la simple application d'une méthode d'alignement des phrases. Nous avons donc

essayé de réaliser un système supportant ces différences structurales des langues traitées. Nous décrivons, dans le présent article, deux méthodes d'alignement des propositions : l'une basée sur les méthodes d'appariement des graphes et une autre inspirée de la classification ascendante hiérarchique (CAH). Nous allons d'abord présenter un bref état de l'art afin de mettre en relief les problèmes. Puis, l'exposé se poursuivra par la description des deux méthodes, pour se terminer par l'analyse des résultats obtenus et une discussion sur les pistes d'amélioration.

2 Problèmes et solution adoptée

Il existe très peu de travaux sur l'alignement des propositions. Nous pouvons tout de même citer ceux de Piperidis, Papageorgiou et Boutsis (Boutsis & Piperidis, 1998) (Piperidis *et al.*, 2000) sur les textes parallèles anglais-grec et ceux de Wang et Ren (Wang & Ren, 2005) sur le japonais-chinois. Dans la méthode de Piperidis, l'algorithme d'alignement est semblable à celui proposé par Brown, Lai et Mercer (Brown *et al.*, 1991) pour l'alignement des phrases, à l'exception du fait qu'il utilise des informations lexicales contrairement à la méthode d'alignement des phrases n'exploitant, elle, que la longueur des textes. Wang et Ren améliorent également l'appariement basé sur les longueurs des textes par l'introduction d'un calcul de similarité basé sur l'information portée par les idéogrammes Han.

Il n'existe à ce jour a priori aucune étude sur l'alignement des propositions traitant le japonais, avec le français, ou même avec l'anglais. Il existe cependant un article portant sur l'alignement manuel des propositions anglais-japonais qui présente une méthode d'alignement manuel et les difficultés rencontrées (Kashioka *et al.*, 2003).

2.1 Difficultés d'appariement des propositions dues aux différences entre les langues

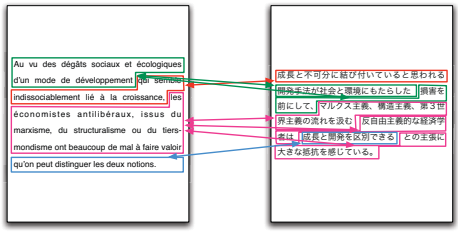


FIG. 1 – Exemple de non-parallélisme de l'alignement des propositions Français-Japonais

Dans leur article, Kashioka *et al.* présentent la constitution d'un corpus parallèle avec alignement au niveau des propositions, réalisée dans un but d'utilisation pour la traduction automatique des monologues (e.g. nouvelles télévisées, conférences, présentations techniques). Une remarque intéressante faite par les auteurs suite à cette expérience, porte sur la différence d'ordre des propositions japonaises et des segments anglais correspondants : on constate beaucoup de

croisement des alignements. Ce non-parallélisme des propositions (cf. figure 1) implique l'impossibilité d'appliquer les méthodes d'alignement des phrases classiques basées sur l'hypothèse de parallélisme. Nous avons donc besoin, pour automatiser la tâche d'alignement des propositions, de concevoir un autre algorithme qui ne présuppose pas le parallélisme.

2.2 Éléments de solution

Pour supporter les croisements des traductions, l'automatisation de l'opération d'alignement nécessite un algorithme utilisant une structure non linéaire mais à deux dimensions telle que les graphes. Notre idée est comme suit : à l'aide des informations sur les relations entre les propositions, nous construisons l'arbre dépendanciel en propositions (arbre des propositions, ci-après) pour réaliser l'alignement par une méthode d'appariement des graphes (cf. figure 2).

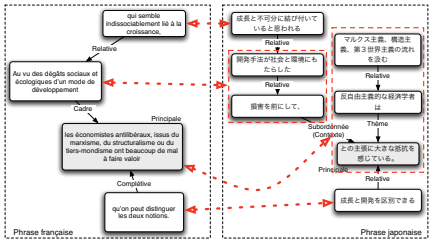


FIG. 2 – Alignement des propositions à l'aide des arbres des propositions

L'alignement à l'aide de graphes n'est pas un concept nouveau. Plusieurs études sur l'alignement anglais-japonais de structures inférieures à la proposition utilisant les arbres syntaxiques ont été réalisées. Les travaux de Matsumoto et *al.* (Matsumoto *et al.*, 1993) proposent une méthode permettant de trouver des correspondances structurelles entre deux arbres de dépendance. Dans les méthodes (Kaji *et al.*, 1993) (Imamura, 2000) (Watanabe *et al.*, 2000), l'alignement des syntagmes est réalisé sur la base des mots mis en correspondance à l'aide d'un dictionnaire bilingue. Les mots alignés servent à ancrer les textes pour repérer les segments à extraire et la représentation arborescente permet de déterminer correctement les structures formées par ces mots ancres.

Notre approche est semblable à celle de Matsumoto. La difficulté est que la recherche de la meilleure décomposition des arbres pour obtenir les structures isomorphes permettant l'appariement maximal revient à un appariement *many-to-many* de graphes, qui est un problème de grande complexité algorithmique. Dans les travaux de Matsumoto, est retenue une stratégie d'amélioration par l'utilisation de la méthode du *branch-and-bound*. Dans le cadre des présents travaux, nous avons choisi une solution basée sur une technique d'appariement des graphes. En effet, dans la théorie des graphes, il existe un ensemble de méthodes beaucoup plus économiques que les procédures de recherche combinatoire, généralement connues sous le nom de méthodes spectrales.

Néanmoins, cette méthode s'appuie essentiellement sur la topologie des graphes à apparier et n'est pas destinée à exploiter différentes informations disponibles, notamment les informations lexicales dans le cas de nos travaux. La dernière étape de la méthode spectrale consistant en un

regroupement des points projetés, nous a inspiré l'approche pour l'alignement par la classification ascendante hiérarchique (CAH). Celle-ci devant permettre de mieux profiter des informations lexicales tout en supportant les croisements des traductions.

Après examen de l'existant, nous avons réalisé deux méthodes d'alignement des propositions. L'une est basée sur les méthodes d'appariement des graphes – profitant pleinement des structures des arbres des propositions –, l'autre exploitant les informations lexicales et les longueurs tout en supportant les croisements de correspondance avec une méthode inspirée de la classification ascendante hiérarchique.

3 Méthodes basées sur l'approche spectrale

Dans la théorie des graphes, l'appariement par une approche spectrale consiste à représenter et distinguer les propriétés structurales des graphes à l'aide des valeurs propres et des vecteurs propres de leurs matrices d'adjacence et se base généralement sur une technique de décomposition spectrale. L'algorithme sur lequel nous nous sommes plus particulièrement appuyés, celui proposé par Kosinov et Caelli (Kosinov & Caelli, 2002) (Kosinov & Caelli, 2004), est une amélioration visant en particulier la réalisation des appariements de graphes inexacts. Lerallut (Lerallut, 2006) a ensuite amélioré cette méthode pour prendre en compte des informations supplémentaires en cas d'appariement de graphes valués.

Dans le cadre de notre alignement des propositions, la méthode de Kosinov est directement utilisée pour apparier les arbres des propositions. Afin d'exploiter au mieux les informations disponibles pour réaliser un meilleur appariement, nous avons également réalisé une adaptation de la méthode de Lerallut à notre opération d'alignement des propositions.

3.1 Méthodes spectrales pour l'appariement des graphes

La méthode d'appariement de graphes inexact proposée par Kosinov et Caelli combine les avantages des techniques de décomposition spectrale, de projection et de classification (*clustering*). Elle consiste, étant donné les matrices d'adjacence A_1 et A_2 créées à partir des graphes G_1 et G_2 respectivement, (i) à calculer les valeurs propres et les vecteurs propres, (ii) à tronquer les matrices selon le nombre de dimensions choisi pour la projection, (iii) à normaliser les valeurs propres et les vecteurs propres pour projeter ensuite chaque graphe, enfin (iv) à réaliser l'appariement par regroupement des nœuds projetés à l'aide d'un algorithme de classification.

Après la décomposition, les données relatives aux nœuds obtenues avec la matrice d'adjacence sont projetées sur les k vecteurs propres les plus importants, formant un sous-espace propre de dimension réduite du graphe. Dans ce sous-espace propre, des nœuds ou des ensembles de nœuds ayant des propriétés structurales semblables sont proches les uns des autres, permettant ainsi une comparaison et un appariement des graphes. Néanmoins, étant donné que les graphes à aligner peuvent posséder un nombre différent de nœuds, une opération de normalisation est également nécessaire pour assurer de bonnes conditions de comparaison. Par examen du positionnement de ces projections de nœuds, la mise en correspondance est alors possible. Le regroupement des points projetés par une méthode de classification ascendante hiérarchique permet de réaliser l'appariement des ensembles de nœuds entre les graphes.

Lerallut, cherchant à appliquer cette méthode à un traitement des images, propose une amé-

lioration permettant de prendre en compte des informations supplémentaires en cas d'appariement de graphes valués. Sa méthode part du résultat obtenu avec la méthode de Kosinov. Cette dernière permet d'abord d'obtenir la matrice topologique contenant des distances euclidiennes entre toutes les projections dans le sous-espace propre. Les graphes sont ensuite valués par l'affectation de couleurs à chaque nœud, et une matrice des distances de couleurs est calculée entre tous les nœuds des deux graphes. Après avoir normalisé ces deux matrices en les divisant par leur valeur maximum, on calcule une somme pondérée. Enfin, après une modification pour écarter les valeurs très distantes, un sous-espace propre de cette matrice est calculé afin d'y projeter tous les nœuds.

3.2 Application de la méthode spectrale à l'alignement des propositions

L'alignement des propositions réalisé par la méthode de Kosinov s'appuie uniquement sur la topologie des graphes. Toutefois, les arbres des propositions dont nous disposons comme entrée du système contiennent beaucoup plus d'informations qui pourraient être utilisées au profit d'un bon appariement. Afin d'exploiter au mieux ces informations disponibles, nous avons tout d'abord tenté d'adapter la méthode de Lerallut de sorte que les graphes à apparier soient valués, non par l'affectation de couleurs, mais selon les types de propositions. Mais, afin de calculer la distance entre deux nœuds sur la base de leur type de proposition, il nous a d'abord fallu définir une distance entre chaque type de proposition.

Faute de corpus *ad. hoc* en quantité suffisante pour le calcul des probabilités des correspondances entre les types, nous avons choisi une méthode plus empirique, qui permettra également de constituer un premier corpus pour des travaux futurs. Nous avons d'abord mis en correspondance les types de propositions du français et du japonais, qui semblaient les plus proches sur le plan syntaxique. Étant donné l'existence d'un lien non négligeable entre les fonctions syntaxiques et la place dans la phrase, nous avons défini une distance entre chaque type de proposition sur la base de la topologie de la phrase. À cette fin, nous nous sommes principalement appuyés sur la structure canonique de la phrase française. La structure canonique est définie comme : éléments extra-prédicatifs – thème – racine (principale) – subordonnées post-nominales – subordonnées post-verbales – subordonnées périphériques[1]. Le principe de base est que la distance d'un type donné de proposition par rapport à la racine, point central, est définie par le nombre de propositions susceptibles d'apparaître entre elles.

Nous avons utilisé les distances ainsi définies pour calculer la matrice des distances de couleurs et réalisé l'appariement des graphes avec la méthode de Lerallut. Mais, nous n'avons pas obtenu le résultat souhaité : l'appariement ne reflétait pas bien les distances des types de propositions. Afin de mieux refléter les informations sur les types de nœuds tout en conservant la structure des arbres d'entrée, nous avons introduit une autre formule. Le principe du nouveau calcul consiste à prendre compte des informations topologiques pour les relations entre les nœuds du même arbre et des informations sur les types pour les distances entre les nœuds d'arbres différents.

Étant donné les deux graphes X et Y, la matrice finale de la méthode de Lerallut est une matrice M_{final} de $|X|+|Y| \times |X|+|Y|$, $M_{\mathit{final}}(i,j)$ correspondant à la somme des distances topologique et de couleur normalisées entre les nœuds i et j.

[1]Pour une description détaillée des types de propositions, voir (Nakamura-Delloye, 2007).

Nous décomposons cette matrice finale comme :

$$M_{\mathit{final}} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

de manière à obtenir les sous-matrices M_{11} comme une matrice $|X| \times |X|$, M_{12} comme $|X| \times |Y|$, M_{21} comme $|Y| \times |X|$ et M_{22} comme $|Y| \times |Y|$, où :

$$\begin{aligned} M_{11}(i,j) &= \mathrm{dist}_{\mathrm{topo}}(X_i, X_j) \times (1-\alpha) \\ M_{12}(i,j) &= \mathrm{dist}_{\mathrm{type}}(X_i, Y_j) \times \alpha \\ M_{21}(i,j) &= \mathrm{dist}_{\mathrm{topo}}(X_j, Y_i) \times (1-\alpha) \\ M_{22}(i,j) &= \mathrm{dist}_{\mathrm{type}}(Y_i, Y_j) \times \alpha \end{aligned}$$

4 Méthode inspirée de la classification ascendante hiérarchique

La deuxième méthode que nous avons décidé d'étudier est basée sur la classification ascendante hiérarchique (CAH), celle-ci devant permettre de mieux profiter des informations lexicales tout en supportant les croisements des traductions. En effet, nous considérons maintenant l'alignement, comme nous l'avons fait dans la méthode spectrale, comme le regroupement des points semblables appartenant à deux classes différentes.

4.1 Deux matrices de base

Nous créons tout d'abord deux matrices : une contenant les similarités de chaque paire de propositions (M_{sim}), et une autre pour stocker les valeurs indiquant l'évolution du rapport des longueurs entre les propositions de langues différentes (M_{raplong}).

Étant donné les deux (ensembles de) phrases X et Y, la matrice de similarité M_{sim} de $(|X| + |Y|) \times (|X| + |Y|)$ est définie comme :

$$M_{\mathit{sim}} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

avec les sous-matrices M_{11} de $|X| \times |X|$, M_{12} de $|X| \times |Y|$, M_{21} de $|Y| \times |X|$ et M_{22} de $|Y| \times |Y|$, où :

$$\begin{aligned} M_{11}(i,j) &= \mathit{synt}(X_i, X_j) \\ M_{12}(i,j) &= \mathit{simlex}(X_i, Y_j) \\ M_{21}(i,j) &= \mathit{simlex}(X_j, Y_i) \\ M_{22}(i,j) &= \mathit{synt}(Y_i, Y_j) \end{aligned}$$

$\mathit{simlex}(X_i, Y_j)$ est la similarité lexicale obtenue de manière classique telle qu'avec le coefficient de Dice. Dans notre réalisation, elle est calculée à l'aide d'un dictionnaire bilingue et d'une liste de mots alignés au moment de l'alignement des phrases du même corpus. Lorsque la similarité lexicale est à 0, on lui donne la valeur minimum α pour favoriser la fusion des éléments (propositions) appartenant aux classes différentes. $\mathit{synt}(X_i, X_j)$ est obtenue de la même manière qu'une matrice d'adjacence, c'est-à-dire 0 s'il n'existe aucun arc entre les nœuds i et j dans l'arbre d'entrée, et β s'il en existe un. Ce mécanisme permet, dans le cas du regroupement d'éléments appartenant à la même classe, de réaliser l'agrégation entre deux éléments en relation de dépendance, plutôt qu'entre deux éléments qui n'en ont aucune.

La matrice d'évolution du rapport des longueurs $M_{raplong}$ est définie telle qu'à chacun de ses éléments (i, j) corresponde l'évolution du rapport des longueurs entre les propositions de langues différentes, qui se produira si le regroupement des deux éléments considérés, i et j, a lieu. La valeur indiquant cette évolution est pondérée par a afin de pénaliser les fusions importantes d'éléments.

$$M_{raplong}(i,j) = \left(rap\left(F(i,j), J(i,j)\right) - \min\left(rap\left(F(i), J(i)\right), rap\left(F(j), J(j)\right)\right)\right) \cdot a$$

où

- $rap(x, y)$ est le rapport des longueurs des éléments (ou des classes) x et y ;
- $F(x)$ (resp. $J(x)$) est la longueur normalisée de (l'ensemble des) proposition(s) française(s) (resp. japonaise(s)), constituant l'élément (ou la classe) x ;
- $F(x, y)$ (resp. $J(x, y)$) est la longueur normalisée de l'ensemble des propositions françaises (resp. japonaises) constituant la classe regroupant les éléments (ou les classes) x et y ;
- a est le poids défini comme le logarithme de la moyenne des deux longueurs normalisées $a = \log\left(\frac{F(i,j)+J(i,j)}{2}\right)$.

4.2 Agrégation et recalcul des matrices

En combinant ces deux matrices, de similarité et d'évolution du rapport des longueurs, une troisième matrice, matrice courante, est calculée et recalculée après chaque agrégation de deux éléments. La matrice courante est définie comme :

$$M_{courante}(i,j) = \frac{M_{raplong}(i,j)}{M_{sim}(i,j)}$$

Dans cette matrice courante, nous cherchons la valeur minimum pour réaliser l'agrégation de deux éléments (ou classes). Après l'agrégation des deux éléments, nous recalculons la matrice d'évolution du rapport des longueurs, en tenant compte du changement de longueurs des éléments regroupés. La matrice de similarité est également recalculée selon le critère d'agrégation adopté. Dans notre réalisation, les similarités des classes nouvellement créées suite à l'agrégation sont obtenues en divisant la somme des similarités des éléments (ou classes) regroupés, par la valeur v calculée sur la base du nombre de propositions faisant partie de cette nouvelle classe. À partir de ces deux matrices nouvellement calculées, on calcule à nouveau la matrice courante et recommençons les opérations d'agrégation tout comme la CAH. À la différence de l'algorithme classique de CAH, l'itération s'arrête dans notre opération dès que toutes les propositions sont regroupées avec au moins une proposition de l'autre langue.

5 Évaluation des méthodes

Nous avons réalisé une évaluation des méthodes proposées avec quatre corpus parallèles de natures diverses et de langues originaires différentes (1, 2 en français et 3, 4 en japonais) : (1) corpus LMD constitué d'articles du Monde Diplomatique, (2) corpus BRVF et (3) BRVJ composés respectivement de deux et d'un brevets techniques, (4) corpus FdT, un extrait du roman « La fin des temps » de Haruki MURAKAMI. Le corpus est d'abord aligné au niveau des phrases par notre système d'alignement des phrases (Nakamura-Delloye, 2005) et le résultat est vérifié manuellement. Puis, pour chaque phrase, la détection des propositions est réalisée à

l'aide de nos détecteurs de propositions du français (Nakamura-Delloye, 2006) et du japonais, et le résultat d'analyse est également corrigé manuellement.

	Caractéristiques				Résultat								
	(A/B)	(C)	(D)	(E)	Partiel (F)			Exact (G)			Paires créées (H)		
	Phr.	Fr	Jp	Prop.	M1	M2	M3	M1	M2	M3	M1	M2	M3
LMD	222/500	644	1026	583	0,643	0,784	0,951	0,127	0,200	0,591	0,813	0,746	0,918
BRVF	161/339	447	854	444	0,619	0,705	0,977	0,081	0,158	0,706	0,750	0,757	0,867
BRVJ	44/66	146	280	141	0,663	0,689	0,990	0,048	0,078	0,537	0,738	0,638	0,674
FdT	99/200	286	428	251	0,670	0,659	0,932	0,138	0,151	0,464	0,892	0,817	0,936

TAB. 1 – Résultats de l'alignement par les trois méthodes

Nous avons utilisé au total 1105 paires de phrases alignées (détails pour chaque corpus indiqués en (B), tableau 1). Parmi celles-ci, nous n'avons pris en compte dans nos résultats d'évaluation que les paires comportant plus d'une proposition dans chaque langue, soit 526 paires de phrases (A), qui représentent 1523 propositions françaises (C) et 2588 propositions japonaises (D), composant 1419 paires de propositions en relation de traduction (E). Nous pouvons constater que le nombre de propositions japonaises est au moins 50% plus élevé que celui des propositions françaises. Cela implique que le modèle de traduction 1-1 (modèle pour la paire en relation de traduction constituée d'une unité dans une langue avec une unité de l'autre langue) est beaucoup moins courant que dans le cas de l'alignement des phrases. En effet, les paires 1-1 représentent moins de 50% et le nombre d'alignements d'une proposition française avec de 2 à plus de 4 propositions japonaises s'élève à environ 40%. Ce type de paire complexe est une source de perturbation pour les méthodes d'alignement des phrases classiques.

Dans le tableau 1, est présenté le résultat de notre évaluation des trois méthodes : méthode spectrale uniquement topologique (M1), méthode spectrale avec types de propositions (M2) et méthodes avec classification ascendante hiérarchique (M3). La zone marquée « Partiel » (F) indique la proportion de paires partiellement correctes parmi l'ensemble des paires effectivement alignées, et la zone marquée « Exact » (G), celle des paires exactement alignées correctement. Enfin, la zone marquée « Paires créées » (H), correspond à la proportion du nombre de paires créées par rapport au nombre correct de paires.

Les résultats montrent que nous avons réussi à améliorer la méthode de Kosinov (M1) avec l'introduction des types de proposition (M2), encore que les chiffres obtenus ne soit pas encore satisfaisants. En effet, il existe beaucoup de cas où la topologie des graphes n'est pas suffisante pour l'appariement des arbres syntaxiques : il arrive, notamment, qu'un arbre soit interprété comme symétrique, alors qu'il ne l'est pas. C'est par exemple le cas des arbres des propositions (présentés figure 3) des phrases parallèles suivantes :

> **Phrase française :** F1 : RACINE Paris avait estimé, à l'époque ‖ F2 : SUBQ , qu'une référence aux valeurs religieuses n'était pas acceptable ‖ F3 : SUBP car elle soulevait des problèmes politiques et constitutionnels en France.
>
> **Phrase japonaise :** J1 : EXTRA *tôji,*(à l'époque) ‖ J2 : THEME *furansu wa,* (La France) ‖ J3 : THEME *shûkyôteki kachi eno genkyû wa* (une référence aux valeurs religieuses) ‖ J4 : SUBAGG *kokunai de seijijô, kenpôjô no mondai wo hikiokosuga yueni* (car [elle] soulève des problèmes politiques et constitutionnels dans le pays) ‖ J5 : SUBCIT *mitomerarenai tono* (qui dit que ce n'était pas acceptable) ‖ J6 : RACINE *shisei wo totta.*([La France] a pris la position)

Dans les cas comme cet exemple, l'introduction des informations sur le type de proposition a permis d'améliorer le résultat et de fournir un alignement correct.

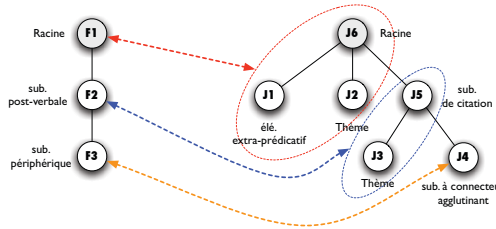


FIG. 3 – Arbres des propositions et appariement correct de leurs nœuds

Toutefois, beaucoup de phrases nécessitaient encore plus d'informations et leur alignement n'a été amélioré qu'avec la méthode à classification ascendante hiérarchique (M3) basée sur la similarité lexicale. Cette méthode possède encore des points potentiels d'amélioration (comme la désambiguïsation lexicale par exemple), mais la capacité d'alignement avec des croisements est un atout crucial. De plus, comme nous le savons bien, la méthode de classification nous permet de définir nous-même la fin du développement des fusions. Par ce mécanisme, nous pourrions obtenir un résultat moins robuste mais plus fiable.

6 Conclusion et perspectives des travaux futurs

Nous avons présenté deux méthodes pour l'alignement des propositions des textes parallèles français-japonais. L'une s'appuie sur une méthode d'appariement des graphes consistant à projeter les nœuds sur un sous-espace propre. L'autre est basée sur une méthode inspirée de la classification ascendante hiérarchique.

Les deux méthodes sont caractérisées par leur capacité d'alignement des traductions croisées dans l'ordre d'apparition, ce qui était impossible pour beaucoup de méthodes classiques d'alignement des phrases. Le résultat obtenu avec la méthode spectrale n'était pas satisfaisant. Il est en effet difficile de trouver une formule permettant de refléter les informations supplémentaires autres que la topologie. Une très récente étude (Fraikin *et al.*, 2006) propose une amélioration visant le traitement des graphes orientés. Néanmoins, du fait des différences considérables de structures, l'application de cette méthode à l'alignement des langues très différentes semble difficile. En revanche, l'alignement basé sur la méthode de classification ascendante hiérarchique est prometteur dans la mesure où cette technique permet d'exploiter plus efficacement différentes informations tout en supportant les croisement des traductions.

À travers cette expérience, nous avons également rencontré beaucoup de constructions pour lesquelles un appariement même manuel était très difficile. Ces exemples sont, pour nous, non seulement des indicateurs de futurs obstacles à franchir, mais aussi très enrichissants du point de vue de l'étude contrastive sur les structures syntaxiques des phrases française et japonaise. Nos travaux sur l'alignement pourraient non seulement participer au développement du domaine du TAL, mais aussi contribuer aux progrès des études linguistiques contrastives du français-japonais, qui favoriseraient à leur tour l'innovation de nos recherches.

Références

BOUTSIS S. & PIPERIDIS S. (1998). Aligning clauses in parallel texts. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, p. 17 – 26.

BROWN P. F., LAI J. C. & MERCER R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, p. 169 – 176.

FRAIKIN C., NESTEROV Y. & VAN DOOREN P. (2006). A gradient-type algorithm optimizing the coupling between matrices and application to graph matching. In *Proceedings of the 13-th ILAS conference in Amsterdam*.

IMAMURA K. (2000). A hierarchical phrase alignment from english and japanese bilingual text. In *Proceedings of CICLing 2001*.

KAJI H., KIDA Y. & MORIMOTO Y. (1993). Learning translation templates from bilingual text. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, p. 672–678.

KASHIOKA H., MARUYAMA T. & TANAKA H. (2003). Building a parallel corpus for monologue with clause alignment. In *Proceedings of the 9th Machine Translation Summit*, p. 216 – 223.

KOSINOV S. & CAELLI T. (2002). Inexact multisubgraph matching using graph eigenspace and clustering models. In *Proceedings of SSPR/SPR*, volume 2396, p. 133–142.

KOSINOV S. & CAELLI T. (2004). An eigenspace projection clustering method for inexact graph matching. *IEEE transactions on pattern analysis and machine intelligence*, **26**(4), 515–519.

LERALLUT R. (2006). *Modélisation et interprétation d'images à l'aide de graphes*. Thèse de doctorat, École des Mines de Paris.

MATSUMOTO Y., ISHIMOTO H. & UTSURO T. (1993). Structural matching of parallel texts. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, p. 23– 30.

NAKAMURA-DELLOYE Y. (2005). Système AlALeR : Alignement au niveau phrastique des textes parallèles français-japonais. In *RECITAL 2005*.

NAKAMURA-DELLOYE Y. (2006). Détection automatique des propositions syntaxiques du français. In *TALN 2006*.

NAKAMURA-DELLOYE Y. (2007). Typologie des subordonnées et des connecteurs en vue de la détection automatique des propositions syntaxiques du français. In *Description linguistique pour le traitement automatique du français*, Cahiers du Cental. Presses universitaires de Louvain. (à paraître).

PIPERIDIS S., PAPAGEORGIOU H. & BOUTSIS S. (2000). From sentences to words and clauses. In J. VÉRONIS, Ed., *Parallel text processing*, p. 117 – 138. Kluwer Academic Publishers.

WANG X. & REN F. (2005). Chinese-japanese clause alignment. *Lecture Notes in Computer Science*, **3406**, 400–412.

WATANABE H., KUROHASHI S. & ARAMAKI E. (2000). Finding structural correspondences from bilingual parsel corpus for corpus-based traslation. In *Proceedings of COLING 2000*, p. 906–912.

Un Lexique Génératif de référence pour le français

Fiammetta NAMER[1], Pierrette BOUILLON[2], Evelyne JACQUEY[3]
[1]Université Nancy2 et ATILF
fiammetta.namer@univ-nancy2.fr
[2] ISSCO
pierrette.bouillon@issco.unige.ch
[3]ATILF,
evelyne.jacquey@atilf.fr

Résumé. Cet article propose une approche originale visant la construction d'un lexique sémantique de référence sur le français. Sa principale caractéristique est de pouvoir s'appuyer sur les propriétés morphologiques des lexèmes. La méthode combine en effet des résultats d'analyse morphologique (Namer, 2002;2003), à partir de ressources lexicales de grande taille (nomenclatures du TLF) et des méthodologies d'acquisition d'information lexicale déjà éprouvées (Namer 2005; Sébillot 2002). Le format de représentation choisi, dans le cadre du Lexique Génératif, se distingue par ses propriétés d'expressivité et d'économie. Cette approche permet donc d'envisager la construction d'un lexique de référence sur le français caractérisé par une forte homogénéité tout en garantissant une couverture large, tant du point de vue de la nomenclature que du point de vue des contenus sémantiques. Une première validation de la méthode fournit une projection quantitative et qualitative des résultats attendus.

Abstract. This paper describes an original approach aiming at building a reference semantic lexicon for French. Its main characteristic is that of being able to rely on morphological properties. The method thus combines morphological analyses results (Namer 2002;2003;2005) from large scale lexical resources (i.e. TLF word lists) with already tested acquisition methodologies on lexical information (Sébillot, 2002). The representation format, within the Generative Lexicon framework, has been chosen for its expressiveness and economy features. So, this approach allows us to consider building a reference lexicon for French, which is fundamentally homogeneous as well as of a large coverage. A feasability study of the described method provides a projection of expected results, from both quantitative and qualitative points of view.

Mots-clés : acquisition lexicale, lexique de référence du français, modèle du lexique génératif, morphologie constructionnelle, corpus, sémantique.

Keywords: lexical acquisition, reference lexicon for French, generative lexicon model, word formation, corpora, semantics.

1 Introduction : objectifs

Cet article présente une méthodologie[1], dont l'objectif est de construire automatiquement un lexique du français dans le format du Lexique Génératif (Pustejovsky 1995), désormais LG. La construction de ce lexique exploite deux sources d'acquisition : des règles linguistiques fondées sur des contraintes imposées par la morphologie (désormais règles de construction de lexèmes : RCL, cf. (Aronoff 1994; Fradin 2003)), et des règles d'apprentissage, à partir de définitions du TLFi[2] ou/et à partir de corpus. Ces dernières s'appliquent aux lexèmes non construits, c'est-à-dire non accessibles aux RCL, et permettent, en cas d'ambiguïté non soluble par les règles, d'acquérir des informations spécifiques, ou d'enrichir les schémas sous-spécifiés. Par cette combinaison d'approches, nous espérons arriver à une meilleure cohérence globale du lexique dérivé. En effet, les RCL définissent des structures sémantiques générales qui s'appliquent de façon systématique à des classes de lexèmes construits similaires. C'est pourquoi elles sont traduisibles sous forme de modèles lexicaux dans le LG. Par exemple, tous les adjectifs déverbaux dénotent une propriété attendue, qui se manifeste sous la forme de l'activation potentielle du prédicat de base. Cette propriété commune s'observe quelles que soient les caractéristiques du verbe de base, comme l'illustrent les exemples sous (1a) (le verbe est agentif) et (1b) (le verbe est inaccusatif) :

(1) a. [[pull lavable]] = λy, $\exists$e, $\exists$x ($\Diamond$ laver'(e, x,y) & pull'(y))
b. [[marchandise périssable]] = λy, $\exists$e ($\Diamond$ périr'(e, y) & marchandise'(y))

Dans la suite, nous présentons cette méthodologie en détail. Nous la situons d'abord dans un cadre plus général, pour montrer sa spécificité. Nous l'illustrons ensuite par plusieurs exemples concrets qui montrent l'apport respectif des RCL et des ressources lexicales et textuelles. Enfin, nous évaluons quantitativement les résultats escomptés.

2 Cadre Théorique

Aujourd'hui, il existe différents lexiques sémantiques pour le français, plus ou moins librement disponibles, par exemple l'adaptation et/ou l'extension en français de ressources comme EuroWordnet (Vossem 2001). Cependant, aucun d'entre eux n'a réussi à rendre explicite l'ensemble des propriétés suivantes : les liens morphologiques entre les entrées lexicales (comme ceux tissés en (1) entre LAVABLE et LAVER, PÉRISSABLE et PÉRIR) ; d'autres relations permettant d'étoffer la structure hiérarchique du lexique (par exemple, le lien nom-verbe entre COUTEAU et sa fonction prototypique, *i.e.* COUPER) ; la polysémie systématique de certains lexèmes (cf. exemples (2)) ; les interactions syntaxe-sémantique (comme la relation mise en jeu, en (2a), entre les deux emplois de COULER et les fonctions grammaticales réalisées dans chaque cas). Or c'est l'acquisition de ces propriétés qui constitue la motivation fondamentale de notre approche.

(2) (a) le bateau coule versus les pirates coulent le bateau
COULER : est-il transitif/causatif ou intransitif/inchoatif ?

[1] développée dans le cadre d'une proposition de projet qui réunit les auteurs de cet article ainsi que C. Fabre (ERSS), P. Sébillot et V. Claveau (IRISA).

[2] TLFi : "Trésor de la Langue Française informatisée": version informatisée du 'Trésor de la langue Française', 16 volumes, cf. URL : http://atilf.atilf.fr/tlf.htm

(b) *commencer* à écrire/lire un livre versus *commencer* un livre
COMMENCER : sélectionne-t-il un objet direct nominal ou un syntagme verbal ?
(c) la salle a applaudi, la salle contient 200 personnes, la salle a été repeinte récemment
SALLE : désigne-t-il un collectif, un espace ou un objet physique ?
(d) *un chien* rapide versus *une route* rapide
RAPIDE : qualifie-t-il la nature (CHIEN) ou la fonction (ROUTE) du nom modifié ?
(e) un livre *rouge* versus un livre *intéressant*
LIVRE : dénote-t-il un objet physique ou de l'information ?

Son originalité tient de fait en deux points : pour extraire ces différentes informations, nous réutilisons un analyseur morphologique existant (DériF, cf. (Namer 2002, 2003, 2005)). Nous tirons ainsi parti des propriétés sémantiques des lexèmes reliés morphologiquement pour dériver un lexique cohérent et de grande taille à partir des traits morphologiques et sémantiques au moyen desquels l'analyseur Dérif annote automatiquement les lexèmes. Ensuite, le modèle choisi pour l'encodage des informations est le LG. Par rapport aux autres théories lexicales, LG présente différents attraits. Il se caractérise fondamentalement par un principe d'économie relayée par un système de types cohérent et des mécanismes génératifs qui factorisent l'information lexicale sémantique et rendent ainsi compte des cas d'ambiguïtés lexicales, comme celles illustrées sous (2). Dans LG, le sens est représenté sous la forme du produit de quatre rubriques, chacune remplissant un rôle particulier. Ces quatre rôles sémantiques, appelés FORMEL, CONSTITUTIF, AGENTIF, et TELIQUE, constituent la structure des qualia. Ils sont illustrés dans la Figure (1) pour le nom COUTEAU.

ROLE	**Fonction**	**Exemple : M = couteau**
FORMEL	place de M dans la taxinomie	x de type *artefact*
CONSTITUTIF	relations (partie-tout) entre M et ses constituants	Relation = Composant-assemblage , entre x (ie le couteau) et z (de type manche) et entre x et y (de type lame)
AGENTIF	conditions nécessaires (présupposées) à l'existence de M	Evénement (accomplissement) : fabriquer, faisant intervenir un agent u et le résultat x
TELIQUE	décrit la finalité de M	Evénement (accomplissement) : couper, faisant intervenir un agent u', l'instrument x et l'objet à couper : y

Figure 1 : Structure de qualia d'une entrée lexicale M dans le LG

Dans cette structure, chaque rôle définit un prédicat qui relie entre eux différents paramètres, et est caractérisé d'un point de vue événementiel, comme l'illustre la troisième colonne. Les paramètres manipulés par le prédicat sont typés. C'est aussi le cas de la structure de qualia, qui, sous la forme du Paradigme Lexical Conceptuel, organise le lexique en classes sémantiques. Toute structure de qualia peut ainsi être interprétée et reformulée dans différents modèles logiques (Pustejovsky 2001). Par exemple, la structure de qualia de COUTEAU peut être traduite dans la formule (3). Cette propriété générale rend le LG particulièrement adéquat pour le calcul logique (Pustejovsky 2001).

(3) λx [couteau'(x) : formal(x) = λx [artefact'(x)] $\wedge$ constitutif(x) = $\exists y,z$ [composant-assemblage'(y, x) $\wedge$ composant-assemblage'(z,x)] $\wedge$ agentif(x) = $\exists e,u$ [fabriquer'(e, u, x)] $\wedge$ telique(x) = $\lambda e'$ $\exists uy'$ [$\Diamond$ couper'(e', u', y, avec'(x))]

Dans le passé, d'autres projets se sont donné un objectif similaire (notamment Acquilex, cf. (Copestake et al. 1993), Simple (Busa et al. 2001) et Clips, cf. (Calzolari et al. 2003)). Cependant aucun d'entre eux n'a véritablement exploité les liens morphologiques pour en

extraire les informations sémantiques pertinentes à grande échelle, faute d'un analyseur morphologique adéquat. En prenant comme source l'analyse fournie par Derif, nous évitons deux pierres d'achoppement : d'abord, la représentation LG découle ici du procédé morphologique impliqué, ce qui la motive théoriquement. D'autre part, tous les lexèmes construits de la même manière reçoivent une représentation LG similaire, ce qui devrait assurer une meilleure cohérence globale de la ressource. Nous espérons ainsi deux retombées principales sur le plan théorique. D'une part, nous complétons l'apport de l'analyseur DériF en le couplant à une sémantique plus profonde. Rappelons que la contribution de la morphologie à la construction du sens lexical, que DériF formalise, n'est que partielle (cf section 3.3) : une RCL ne fournit en effet que les éléments fondamentaux de l'interprétation d'un lexème, qui seront spécifiés ultérieurement, notamment par le contexte d'utilisation (cf. (Aronoff 1980)). D'autre part, nous confirmons les hypothèses théoriques sur lesquelles repose le LG, et notamment nous répondons à la question suivante : LG fournit-il un cadre suffisant pour décrire les différents aspects de la sémantique lexicale et ceci, à grande échelle, pour différentes familles de lexèmes ? Idéalement, toutes les propriétés prédictibles par les RCL (par exemple : la base verbale des verbes préfixés par *dé-* , e.g. DÉCOUDRE, DÉFAIRE, dénote un *accomplissement*, dont le résultat désigne une propriété *réversible*, cf. (Amiot à paraître)) devraient pouvoir être représentées formellement au niveau de la structure des qualia. De cette structure, différentes informations peuvent ensuite être extraites, dont on a déjà montré l'intérêt sur le plan pratique. (Bouillon et al. 2000; Claveau et al. 2001; Claveau et al. 2003) ont notamment montré comment tirer parti des liens Nom-Verbe exprimés dans la qualia pour la recherche documentaire. Une question ouverte est de savoir si cette ressource sera plus utile que les précédentes basées sur le même formalisme. Nous pensons en tout cas que le fait de lier cette ressource à un analyseur morphologique existant la rend potentiellement plus apte à répondre à la question fondamentale de la créativité de la langue, DériF étant conçu pour l'analyse des lexèmes inconnus. Dans la suite, nous décrivons plus en détail la méthodologie d'acquisition.

3 Méthodologie d'acquisition du lexique

L'acquisition des entrées lexicales fait collaborer deux approches décrites dans ce qui suit. La première se fonde sur l'utilisation de connaissances morphologiques (section 3.1), la seconde sur l'emploi de méthodes d'apprentissage à partir de corpus (section 3.2). La combinaison des deux approches (section 3.3), enfin, confirme, infirme ou affine les contributions propres à chaque technique.

3.1 Acquisition d'entrée par règles morphologiques

Cette approche exploite les résultats fournis par l'analyseur morphologique DériF (Namer 2002, 2003, 2005), qui sont reformatés de manière à être en conformité avec les notations du LG, suivant l'extension du modèle proposé dans (Jacquey et al. à paraître; Namer et al. 2003, soumission).

3.1.1 Résultats de DériF

DériF produit la décomposition d'un lexème L muni d'une catégorie grammaticale, selon sa base morphologique B. Cette analyse s'accompagne d'annotations, reflétant les contraintes de la RCL appliquée, et pouvant porter sur L et B. Par exemple, la Figure2 illustre au moyen de

l'analyse de DESSOULER$_{VERBE}$ (construit sur SOUL$_{ADJ}$) le format dans lequel s'affiche le résultat de l'analyse par la RCL en *dé-* des verbes désadjectivaux. En dehors de l'analyse elle même (lignes 1 et 2), la règle attribue aux lexèmes reliés les traits suivants : l'adjectif doit être toujours **qualificatif** et décrire une propriété **transitoire** (ligne 3). Le verbe est soit **transitif causatif** ("le café salé dessoule Max") soit **intransitif résultatif** (ou anticausatif) ("Max dessoule") (ligne 4).

```
1  dessouler/VERBE ==> soul,ADJ/dé:prefixe
2  "(Supprimer - Faire perdre) le caractère soul"
3  soul/ADJ:(prédicatif,_,temporaire)
4  dessouler/VERBE:  (causatif,transitif,[cause,theme])||
                     (resultatif,intransitif,[theme])
```

Figure 2 : Analyse du verbe DESSOULER par DériF

3.1.2 Traduction des résultats dans le LG

En fonction des résultats produits par DériF, il est au mieux possible de générer automatiquement les deux entrées au format LG (celles du lexème analysé L et de sa base B) reliées morphologiquement par la règle ayant produit le résultat. Le niveau de spécification de chaque entrée dépend des informations produites lors de la phase d'analyse morphologique. Dans l'exemple pris plus haut, il est ainsi possible de prédire la distribution sous forme de rôles de qualia des étapes de l'enchaînement causal qui constituent l'événement complexe défini par le verbe DESSOULER (Figure 3a), à savoir la succession des situations suivantes : (1) l'individu y est soul (état initial, présupposé dans le rôle AGENTIF) ; (2) l'agent x dessoule l'individu y ou y dessoule (accomplissement, dans le rôle AGENTIF) ; (3) y n'est plus soul (état final, rôle FORMEL). En ce qui concerne l'adjectif SOUL, en tant que base de DESSOULER, la seule information inférable à partir de la règle de préfixation en *dé-* est qu'il s'agit d'une propriété identifiée d'un point de vue événementiel sous la forme d'un état transitoire e (ou stage level predicate, cf. (Carlson 1977)) affectant un individu y (cf. Figure 3b) (dans l'ontologie du rôle FORMEL, e caractérise SOUL comme sous-type du type 'état'). La valeur des autres traits (Structure Argumentale, Structure Evénementielle) constituant une entrée lexicale dans le LG est ensuite instanciée à partir du contenu de la Structure de qualia.

ROLE	**(V) dessouler**
FORMEL	not (soul' (e1 :état_trans,y :individu))
AGENTIF	soul'(e0 :etat_trans,y) ET dessouler_acte'(e2 :accompl., x :agent, y) OU soul'(e0 :etat_trans,y) ET dessouler_acte' (e2 :accompl., y)

ROLE	**(A) soul**
FORMEL	soul'(e :état_trans, y :patient)

Figure 3 : Décomposition des sens de DESSOULER (a) et de SOUL (b) sous forme de rôles de qualia, à partir de l'analyse par DériF

3.2 Acquisition par apprentissage sur corpus

En dehors des cas où les lexèmes simples sont des bases d'autres lexèmes construits, DériF ne fournit aucune information à traduire dans le format LG. Dans ces cas de silence, la méthodologie présentée s'appuie sur un apprentissage de corpus, et plus précisément sur un

corpus issu des données lexicographiques du TLFi. Une expérience préliminaire, effectuée sur la version XML catégorisée du TLFi, montre par exemple que les définitions de ce dictionnaire sont suffisamment régulières pour permettre de détecter les substantifs ayant une fonction prototypique, c'est-à-dire une facette télique dans leur contenu sémantique. En recherchant dans le corpus des définitions des expressions comme « servir, permettre, destiné à/au/aux/de », on repère près de 14% de substantifs ayant un emploi télique (4279 des 30544 substantifs du TLFi). Ceci constitue donc un premier résultat non négligeable. De plus, les expressions verbales ci-dessus s'accompagnent de substantifs (« appareil, organe, instrument ») qui permettent d'identifier des rôles formels (« balai = Ustensile$_{FORMEL}$ de ménage servant au **nettoyage**$_{TELIQUE}$ »). Ces substantifs servent à leur tour à la recherche de nouvelles expressions verbales, permettant ainsi de détecter de nouveaux prédicats téliques, et ainsi de suite… Cette première expérience montre donc l'exploitabilité du corpus des définitions catégorisées du TLFi dans le cadre de la construction automatique d'un lexique du français.

3.3 Croisement des deux approches

Le croisement des deux méthodes d'acquisition présentées *supra* peut servir à préciser certaines informations laissées sous-spécifiées par l'application des RCL. La sous-spécification a deux origines. Soit la morphologie ne dispose pas d'indices suffisants pour préciser un sens (cf section 3.3.1), soit le lexème construit est intrinsèquement ambigu (cf. section 3.3.2). Dans les deux cas, la contribution de l'approche par corpus doit servir à apporter l'information manquante.

3.3.1 Morphologie et sous-spécification : la préfixation en dé-

Considérons le cas des verbes en *dé-* sur base nominale. Deux types de sens sont possibles pour ces verbes. En effet, soit le nom de base décrit la localisation initiale de ce que dénote l'objet direct du verbe ; c'est ce que l'on observe avec DÉTERRER : « faire sortir qqch de terre ». Soit, au contraire, le nom de base décrit l'entité qui subit le changement de lieu ; c'est ce que l'on observe avec DÉSOSSER : « faire sortir les os de qqch/quelquepart ». L'analyse automatique par DériF est incapable de distinguer les deux cas de figure, qui se différencient sur la base de caractéristiques des noms de base qui sont d'ordre extralinguistique. Par conséquent, DériF fournit systématiquement deux définitions à chaque analyse de ce type de verbe. Ainsi, pour DÉTERRER, on obtient :

```
1  déterrer/VERBE ==> terre,NOM/dé:prefixe
2  "Faire sortir qqc de terre || Faire sortir la terre de qqc"
3  déterrer/VERBE:  (dynamique,trans. ,[cause,theme], causatif)
```

Figure 4 : Analyse de DETERRER par DériF

Cette analyse illustre un certain nombre de faits. Tout d'abord, aucune contrainte n'est généralisable pour la base. En effet, le nom sélectionné par la règle en *dé-* est soit concret (TERRE : DETERRER), soit abstrait (COURAGE : DÉCOURAGER). Par contre, la règle impose que le verbe résultant (DÉTERRER, DECOURAGER) soit transitif et décrive un accomplissement (plus précisément, un changement de localisation). En résumé, les contraintes liées à la formation de verbes dénominaux en *dé-* sont les suivantes : (1) le verbe est transitif, causatif, faisant intervenir une cause **x** et un thème **y**; (2) le procès désigne l'acte du causateur sur le thème (« x cause qqc à y »); (3) l'état final affecte soit **y** (c'est le cas avec DÉTERRER), soit

l'entité décrite par le nom de base (e.g. avec DÉCLOUER). Avec DÉTERRER c'est **y** qui est une entité délocalisée par rapport à son site[3] initial : TERRE (*Max déterre le coffre*). Par contre, avec DÉCLOUER (ex : *Max décloue le coffre*), **y** joue le rôle du site initial alors que CLOU est la cible (voir note précédente) délocalisée par rapport au site. La RCL est donc incapable de déterminer qui, du thème ou de la base d'un verbe quelconque préfixé par *dé-*, joue le rôle de cible (et de site), et donc, en conséquence, de définir le verbe construit de façon univoque par rapport à sa base. Voilà pourquoi deux définitions sont proposées pour DÉTERRER, comme l'indique la ligne 2 de la Figure 4. Face à ce type d'ambiguïté que des règles linguistiques ne peuvent pas résoudre, deux cas de figures peuvent se présenter : soit un seul des deux sens est attesté, soit les deux emplois existent, mais avec des fréquences différentes. Dans les deux cas, l'analyse de corpus et de définitions du TLFi permet, selon le cas, soit de lever l'ambiguité, soit de pondérer chacun des emplois. Avec 'DÉTERRER X', par exemple, l'interprétation « faire sortir la terre de X » est improbable : elle est absente des définitions du TLFi et des 100 premiers résultats renvoyés par Google.

3.3.2 Morphologie et exceptions : la suffixation en -oir

Alors que l'exemple précédent montre comment les définitions dictionnairiques et les connaissances textuelles précisent l'information lexicale que la morphologie ne sait qu'esquisser, voyons maintenant un cas inverse : l'analyse de corpus au service de la détection des exceptions à une RCL, à savoir la formation de noms déverbaux en *–oir*. A l'exception de quelques noms désignant le patient prototypique du verbe de base (TIROIR), ou son agent (CONSOLOIR [4]), la RCL *–oir* construit des noms faisant référence à des lieux ou des instruments [5]. La différence entre ces deux concepts est parfois ténue, dès lors qu'un objet possède la taille requise pour occuper les deux fonctions : c'est le cas de la majeure partie des noms déverbaux en *–oir* (ABREUVOIR, BALANÇOIRE, ÉGOUTTOIR désignent chacun l'objet qui sert à la fois d'instrument aidant au déroulement du procès décrit par le verbe, et de lieu où ce procès se déroule). D'autres noms en *–oir* ont la particularité d'être polyréférentiels : ils dénotent deux objets distincts, comme HACHOIR qui désigne soit un couteau soit une planchette, ou SOULOIR, qui en argot fait référence à un verre ou à un débit de boisson. Enfin, certains noms identifient clairement un lieu (DORTOIR, FUMOIR) ou un instrument (RATISSOIR, RASOIR). Comme le montre la Figure 5, l'analyse par DériF des noms déverbaux en *–oir* reflète la polysémie qui caractérise la plupart de ces noms. Presque rien, par contre, n'est prédictible pour le verbe de base : il peut être transitif (ABREUVOIR), ergatif (DESSOULOIR), inergatif (TROTTOIR), inaccusatif (MOUROIR). En bref, à quelques exceptions près (on relève CRÊCHOIR et VIVOTOIR sur la Toile, dont le verbe de base est analysable comme statif), la seule propriété verbale identifiable est 'dynamique'.

```
abreuvoir/NOM ==> abreuver,VERBE/oir:suffixe
```

[3] Les termes de cible et site sont empruntés à (Vandeloise 1986). D'autres linguistes du courant cognitiviste utilisent les termes de, respectivement, trajector et landmark (Langacker 1987) ou figure et ground (Talmy 1983).

[4] "Elle est mon refuge, mon consoloir", (Yahoo).

[5] Entre autres définitions des rôles thématiques, celle de (Fillmore 1968) dit qu'un instrument est "la force ou objet inanimé impliqué dans et à l'origine de l'événement" et un lieu désigne "la localisation ou l'orientation spatiale de l'événement"

```
"Instrument de abreuver || Lieu de abreuver"
abreuver/VERBE:     (dynamique,-,-,-)
```

Figure 5 : Analyse de ABREUVOIR par DériF

La représentation dans le LG des propriétés de ABREUVOIR prédites par la RCL *–oir* exprime donc les faits suivants: (1) le nom désigne une entité, qui possède un type pointé (instrument•lieu), mécanisme du LG pour exprimer les différentes facettes des lexèmes polysémiques ; (2) il est l'un des participants du prédicat (dynamique) potentiel décrit par le verbe de base et qui définit le rôle TELIQUE (c'est à dire la fonction prévue de l'entité décrite par le nom). Tout nom construit en *–oir* est codé suivant ces indications, qui révèlent l'ambiguïté supposée par défaut de ce type de noms. Cette hypothèse doit alors être vérifiée dans les corpus : le nom apparaît-il derrière la préposition "avec" ? dans un complément locatif ? En fonction de la réponse, soit la polysémie est confirmée, soit le codage est revu (instrument ou lieu seul) en fonction de l'échec à l'un ou l'autre des tests.

4 Evaluation

Nous sommes dès à présent en mesure de prévoir quantitativement et qualitativement une partie des résultats qui seront produits selon cette approche, à savoir les entrées lexicales générées à partir des analyses de DériF. Actuellement, au moins 35,5% (i.e. 35263/99445) des lexèmes du TLFi sont analysés comme construits par DériF, sous la forme de 45478 étapes constructionnelles[6]. Ce résultat correspond à l'activation d'environ 85 RCLs. Le pourcentage de 35,5% a deux motifs principaux : (1) toutes les RCL ne sont pas encore implémentées, (2) une grande partie des entrées lexicales du TLFi ne sont pas morphologiquement construites. La Figure 6 montre les règles les plus fréquemment appliquées.

Type d'opération de dérivation morphologique	Catégorie du Construit : Règle (Catégorie de la Base)
Suffixation	A :el, ique, if, al,eux, aire, ien, iste (N), **able**(V), N : **eur** (V), ie, **ité** (A), V :**ifier**, **iser**(N,A), re (V)
Préfixation	V :**en**, a (A,N), **é, dé** (A,N,V), pré (V) A :**in**, hyper,sub,non (A), sur,anti,sub,sous, mono, poly, auto (N)
Conversion	N ->V, **A->V,** V->N, A -> N

Figure 6 : Règles s'appliquant à plus de 80 lexèmes du TLFi

Parmi celles-ci, les règles en gras sont d'ores et déjà associées à des contraintes pour l'identification de leur entrée ou de leur sortie prototypique, à l'image de ce qui est déjà illustré par les Figures 2, 4, 5 (cf. *supra*) ainsi que la Figure 7, qui reprend ci-dessous les conditions de formation des adjectifs déverbaux en *–able*, ébauchées dans la section 1, cf. exemple (1).

```
lavable/ADJ ==> laver,VERBE/able:suffixe
  "Que l'on peut laver || PREP que l'on peut laver"
```

[6] Soit 39028 opérations de dérivation, et 6450 opérations de composition dite néoclassique (cf *infra*).

```
laver/VERBE: (dynamique,-,[-,theme], -)
lavable/ADJ :     (prédicatif, latent,-)
```

Figure 7 : Analyse de LAVABLE par DériF

La formation d'adjectifs déverbaux en *–able* illustre d'ailleurs un autre cas de collaboration nécessaire entre approche morphologique et recherche en corpus. En effet, la contrainte par défaut stipule que le nom recteur de l'adjectif s'identifie avec le patient du verbe de base de celui-ci (*laver un pull / un pull lavable*). Cependant, (Hathout et al. 2003) ont montré que selon l'adjectif (et donc son verbe de base) n'importe quel participant du prédicat pouvait occuper cette fonction : « *un poisson, une saison, un étang pêchable* ». Ce n'est donc qu'en retrouvant en corpus la construction verbale (« *pêcher un poisson, **dans** un étang, **pendant** une saison* ») que l'on pourra affiner l'entrée lexicale de l'adjectif et du verbe reliés par la RCLable. Pour finir, rappelons que l'avantage de DériF est qu'il s'applique à n'importe quel lexique, qui peut ainsi servir d'entrée à la méthodologie et accroître la taille de la ressource LG. Par exemple, la substitution d'un lexique spécialisé biomédical à la liste des nomenclatures du TLFi fait varier sensiblement les chiffres présentés dans la Figure 6 : ici, 59% des lexèmes (soit 17297 des 29273 entrées) sont analysés comme construits, essentiellement par composition néoclassique[7] (13237 des 21757 opérations morphologiques analysées). Une évaluation quantitative récente de DériF par rapport à un "Gold Standard" ((Namer et al. 2007)) a montré un score d'au moins 77,6% de bonnes analyses de la part du programme. Une autre expérience ((Namer 2007b),(Namer et al. 2007),(Deléger et al. 2007)) a, elle, prouvé que, appliqué aux lexiques spécialisés du domaine biomédical, les RCL de DériF sont facilement transposables dans d'autres langues; en particulier, des règles d'analyses des composés néoclassiques ont été traduites en anglais avec succès ((Deléger et al. 2007)).

5 Conclusion

Dans cet article, nous avons présenté une méthodologie pour dériver un LG. Son originalité repose surtout sur l'exploitation des propriétés morphologiques du lexique. Sur le plan pratique, cette approche nous permet de garantir la cohérence des informations dérivées ; sur le plan théorique, elle montre comment la morphologie peut collaborer avec d'autres connaissances (dictionnairiques et textuelles) pour dériver des représentations profondes du sens des mots. Nous ne connaissons pas d'autre tentative du même type pour exploiter ensemble l'apport de ces différentes disciplines. Les résultats attendus comprennent le lexique LG et une série d'outils pour extraire dynamiquement de nouvelles entrées, en particulier Dérif, des modèles pour convertir la sortie de Dérif en des modèles LG et des règles d'extraction pour le corpus/TLFi.

Références

AMIOT, D. (à paraître). La catégorie de la base dans la préfixation en *dé-*. *La raison morphologique. Hommage à la mémoire de Danielle Corbin*. B. Fradin. Amsterdam/Philadelphia, John Benjamins.
ARONOFF, M. (1980). Contextuals. *Language* **56**(4): 744-758.
ARONOFF, M. (1994). *Morphology by Itself*. Cambridge, MIT Press.

[7] Un nom ou adjectif composé néoclassique est formé par une règle de composition, et diffère à plusieurs égards des composés dits 'standard' ou 'ordinaires': leur sens, leur structure, les composants impliqués, les domaines textuels et/ou le registre de langue concernés, etc. cf. (Namer 2007a).

BOUILLON, P., C. FABRE, P. SÉBILLOT and L. JACQMIN (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL* **41**(2): 367-393.
BUSA, F., N. CALZOLARI and A. LENCI (2001). Generative Lexicon and the Simple model, developing Semantic Resources for NLP. *the Language of Word Meaning*. P. Bouillon and F. Busa. Cambridge, CUP**:** 333-349.
CALZOLARI, N., F. BERTAGNA, A. LENCI and M. MONACHINI (2003). New perspectives for lexical web resource in the Semantic Web Scenario. *Generative Approaches to the Lexicon*, Geneva.
CARLSON, G. (1977). Reference to Kinds in English, University of California. **Ph.D. Dissertation**.
CLAVEAU, V., P. SÉBILLOT, P. BOUILLON and C. FABRE (2001). Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? *TAL* **42**(3): 729-753.
CLAVEAU, V., P. SÉBILLOT, C. FABRE and P. BOUILLON (2003). Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming. *Journal of Machine Learning Research* **4**: 493-525.
COPESTAKE, A., S. ANTONIO, T. BRISCOE and V. DE PAIVA (1993). The ACQUILEX LKB : an introduction. *Inheritance, defaults and the lexicon*. T. Briscoe, A. Copestake*et al*. Cambridge, CUP**:** 148-163.
DELÉGER, L., F. NAMER and P. ZWEIGENBAUM (2007). Analyse morphosémantique des composés savants : transposition du français à l'anglais. *TALN*, Toulouse.
FILLMORE, C. (1968). The case for case. *Universals in Linguistic Theory*. E. Bach and R. Harms. New-York, Holt, Rinehart, and Winston**:** 1-88.
FRADIN, B. (2003). *Nouvelles approches en morphologie*. Paris, Presses Universitaires de France.
HATHOUT, N., M. PLÉNAT and L. TANGUY (2003). Enquête sur les dérivés en *-able*. *Cahiers de Grammaire*. N. Hathout, M. Rochée*t al*. Toulouse, ERSS. **28:** 49-91.
JACQUEY, E. and F. NAMER (à paraître). Morphosémantique et modélisation : le cas des verbes dénominaux préfixés par *é-*. *Actes du Colloque "Le sens en linguistique" (25-27 mai 2003)*, Montréal.
LANGACKER, R. (1987). *Foundations of Cognitive Grammar*. Standford, Standford University Press.
NAMER, F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. *Actes de Traitement Automatique du Langage Naturel (TALN) 2002*, Nancy, France, ATALA-ATILF.
NAMER, F. (2003). Automatiser l'analyse morpho-sémantique non affixale : le système DériF *Cahiers de Grammaire*. N. Hathout, M. Rochée*t al*. Toulouse, ERSS. **28:** 31-48.
NAMER, F. (2005). *La Morphologie Constructionnelle du Français et les Propriétés Sémantiques du Lexique - Mémoire présenté dans le cadre de l'habilitation à diriger des recherches*. UFR Sciences du Langage. Nancy, Université de Nancy2.
NAMER, F. (2007a). Composition néoclassique : est-on dans l' "hétéromorphosémie" ? *Morphologie à Toulouse - Actes du colloque international de Morphologie 4èmes Décembrettes*. N. Hathout and F. Montermini. München, Lincom Europa. **(LSTL 37):** 185-206.
NAMER, F. (2007b). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. *T.A.L.* **46**(2): 157-181.
NAMER, F. and R. BAUD (2007). Defining and relating biomedical terms : towards a cross-language morphosemantics-based system. *International Journal of Medical Informatics* **76**: 226-233.
NAMER, F. and E. JACQUEY (2003). Lexical Semantics and derivational morphology: the case of the popular *é-* prefixation in French *GL 2003 : 2nd International Workshop on Generative Approaches to the Lexicon (May, 15-17 2003)*, Geneva.
NAMER, F. and E. JACQUEY (à paraître). Word Formation Rules and the Generative Lexicon: Representing noun-to-verb versus verb-to-noun Conversion. *Generative Lexicon Book*. P. Bouillon, H. Isahara*et al*. Dordrecht, Kluwer.
PUSTEJOVSKY, J. (1995). *The Generative Lexicon*. Cambridge, MA, MIT Press.
PUSTEJOVSKY, J. (2001). Type Construction and the Logic of Concepts. *The Syntax of Word Meanings*. P. Bouillon and F. Busa. Cambridge, Cambridge University Press**:** 91-123.
SÉBILLOT, P. (2002). *Apprentissage sur corpus de relations lexicales sémantiques - La linguistique et l'apprentissage au service d'applications du traitement automatique des langues. Mémoire présenté dans le cadre de l'habilitation à diriger des recherches.* Rennes, Université de Rennes 1.
TALMY, L. (1983). How language structures space. *Spatial orientation, Theory, Research, and Applications*. H. Pick and L. Acredolo. New York, Plenum**:** 225-282.
VANDELOISE, C. (1986). *L'espace en français: sémantique des prépositions spatiales* Paris, Les éditions du Seuil.
VOSSEM, P. (2001). Condensed Meaning in EuroWordnet. *The language of Word Meaning*. P. Bouillon and F. Busa. Cambridge, CUP**:** 363-383.

Les résultats de la campagne EASY d'évaluation des analyseurs syntaxiques du français

Patrick PAROUBEK[1], Anne VILNAT[1], Isabelle ROBBA[1], Christelle AYACHE[2]
[1] LIMSI-CNRS Bât. 508 Université Paris XI, BP 133 - 91403 ORSAY Cedex
[2] ELRA-ELDA 55-57, rue Brillat Savarin 75013 Paris
{pap,anne,isabelle}@limsi.frayache@elda.fr

Résumé. Dans cet article, nous présentons les résultats de la campagne d'évaluation EASY des analyseurs syntaxiques du français. EASY a été la toute première campagne d'évaluation comparative des analyseurs syntaxiques du français en mode boîte noire utilisant des mesures objectives quantitatives. EASY fait partie du programme TECHNOLANGUE du Ministère délégué à la Recherche et à l'Éducation, avec le soutien du ministère de délégué à l'industrie et du ministère de la culture et de la communication. Nous exposons tout d'abord la position de la campagne par rapport aux autres projets d'évaluation en analyse syntaxique, puis nous présentos son déroulement, et donnons les résultats des 15 analyseurs participants en fonction des différents types de corpus et des différentes annotations (constituants et relations). Nous proposons ensuite un ensemble de leçons à tirer de cette campagne, en particulier à propos du protocole d'évaluation, de la définition de la segmentation en unités linguistiques, du formalisme et des activités d'annotation, des critères de qualité des données, des annotations et des résultats, et finalement de la notion de référence en analyse syntaxique. Nous concluons en présentant comment les résultats d'EASY se prolongent dans le projet PASSAGE (ANR-06-MDCA-013) qui vient de débuter et dont l'objectif est d'étiqueter un grand corpus par plusieurs analyseurs en les combinant selon des paramètres issus de l'évaluation.

Abstract. In this paper, we present the results of the EASY evaluation campaign on parsers of French. EASY has been the very first black-box comparative evaluation campaign for parsers of French, with objective quantitative performance measures. EASY was part of the TECHNOLANGUE program of the Delegate Ministry of Research, jointly supported by the Delegate Ministry of Industry and the ministry of Culture and Communication. After setting EASY in the context of parsing evaluation and giving an account of the campaign, we present the results obtained by 15 parsers according to syntactic relation and subcorpus genre. Then we propose some lessons to draw from this campaign, in particular about the evaluation protocole, the segmenting into linguistic units, the formalism and the annotation activities, the quality criteria to apply for data, annotations and results and finally about the notion of reference for parsing. We conclude by showing how EASY results extend through the PASSAGE project (ANR-06-MDCA-013), which has just started and whose aim is the automatic annotation of a large corpus by several parsers, the combination of which being parametrized by results stemming from evaluation.

Mots-clés : analyseur syntaxique, évaluation, français.

Keywords: parser, evaluation, french.

1 L'évaluation des analyseurs syntaxiques

Les premières tentatives d'évaluation des analyseurs ont été le fait d'experts qui fondaient leur appréciation d'un analyseur sur les observations qu'ils avaient faites de ses sorties sur différentes phrases de test, parfois aidés d'une grille d'analyse (Blache & Morin, 2003). Pour le français, à notre connaissance la première tentative d'évaluation comparative a été faite par A. Abeillé (Abeillé, 1991). Dans le souci de réduire la part de subjectivité dans le processus d'évaluation et pour réutiliser les connaissances acquises lors d'une évaluation, les chercheurs se sont ensuite tournés vers des jeux de test prédéfinis, dont TSNLP (Oepen *et al.*, 1996), qui contient des exemples d'analyses correctes et erronées classés par type de constructions linguistiques, est un archétype. Cependant les jeux de test ne peuvent pas rendre compte de la distribution des phénomènes dans un corpus. De plus leur utilité à des fin d'évaluation dans des campagnes ouvertes est limitée dès lors qu'ils sont rendus publics. En effet, il sont de petite taille et paramétrer un analyseur en fonction d'un jeu de test donné devient alors une tâche aisée.
Avec le développement conjoint des standards pour les méta-données et des capacités des ordinateurs, nous avons vu appraître les corpus arborés (*treebanks*), dont le plus célèbre est certainement le Penn Treebank (Marcus *et al.*, 1993). Depuis sa création de nombreux développements pour différents formalismes et pour différentes langues ont vu le jour, dont certains pour le français (Brant *et al.*, 2002) (Abeillé *et al.*, 2000). Cependant, si les corpus arborés peuvent apporter un élément de réponse en ce qui concerne la représentativité des différents genres de texte et la distribution des phénomènes linguistiques, ils n'apportent pas de réponse au problème du formalisme pivot, pour lequel il n'existe à ce jour aucun standard [1].

Comparer des analyseurs implique donc de pouvoir projeter leurs annotations dans une représentation unique, ce qui en général ne peut se faire sans perte d'information. Pour résoudre ce problème, certains (Gaizauskas *et al.*, 1998) proposent de définir une fonction entre systèmes d'annotation, d'autres de tenir compte de la quantité d'information (Musillo & Sima'an, 2002) (méthode qui a le désavantage de nécessiter la construction d'un corpus parallèle par formalisme d'annotation), d'autres encore proposent d'utiliser des mécanismes d'apprentissage grammatical ou des mesures basées sur la distance d'édition (Roark, 2002). En remontant un peu plus dans le passé, (Black *et al.*, 1991) fut le premier à proposer une mesure d'évaluation fondée sur les limites des constituants pour comparer les analyseurs en mesurant le taux de croisement des frontières avec les annotations de référence (*crossing brackets*) et le rappel. En ajoutant la précision aux deux mesures précédentes, on obtient le protocole GEIG (Grammar Evaluation Interest Group) (Srinivas *et al.*, 1996), ou mesures PARSEVAL (Carroll *et al.*, 2002). Cependant ces mesures ont été appliquées uniquement sur des constituants non étiquetés, car il était impossible alors de définir un jeu d'étiquettes commun (Black *et al.*, 1991).

À part quelques tentatives ponctuelles, de comparaisons d'analyseurs syntaxiques, comme celle du projet SPARKLE qui a comparé des analyseurs syntaxiques pour déterminer le plus approprié pour une tâche d'extraction terminologique, ou encore les expériences développées récemment sur des transcriptions orales (Roark *et al.*, 2006), le paradigme d'évaluation n'a jusqu'à présent pas été appliqué à l'analyse syntaxique sur une grande échelle, à l'exception du projet EASY (Vilnat *et al.*, 2004) (Paroubek *et al.*, 2005) qui concerne les analyseurs du français.

[1] Une proposition est en cours d'élaboration à l'ISO.

2 La campagne EASY

La campagne EASY était une des 8 campagnes d'évaluation des technologies de la langue du projet EVALDA du programme TECHNOLANGUE (décembre 2002 - avril 2006). Dans cette campagne, 15 analyseurs provenant de 13 participants différents : ERSS, FT R&D, INRIA, LATL, LIC2M, LIRMM, LORIA, LPL, STIM, SYNAPSE, SYSTAL, TAGMATICA, VALORIA et XRCE ont été évalués sur les données fournies par les 5 fournisseurs de corpus que sont l'ATILF, le LLF, le DELIC, le STIM et ELDA. La tâche des fournisseurs de corpus a consisté en la collecte du corpus de différents genres de textes et en leur annotation. Le rapport entre la portion de texte annoté et la taille totale du corpus est choisie de manière à décourager une annotation manuelle de l'intégralité du corpus. Le corpus contient des articles de journaux (*Le Monde*), des textes littéraires (issus de la base *Frantext* de l'ATILF), des textes médicaux (pathologies et traitements), des questions (issues de la campagne EQUER de TECHNOLANGUE), des transcriptions de débats parlementaires (Sénat français et Parlement Européen), des pages WEB du site ELDA, des courriers électroniques et des transcriptions de parole[2]. On pourra trouver dans le tableau 4 plus loin dans l'article, les tailles respectives de ces différents corpus.

Le protocole d'évaluation EASY suppose que tous les participants adoptent la même segmentation en mots et en énoncés (voir (Roark, 2002) pour les problèmes que cela pose). Le formalisme inspiré de (Carroll *et al.*, 2002) et défini en collaboration avec les participants doit permettre d'exprimer l'essentiel d'une annotation syntaxique quelle que soit son type (de surface ou profonde, complète ou partielle), ceci sans privilégier une approche particulière. Le formalisme d'annotation EASY permet d'annoter des constituants continus et non-récursifs ainsi que des relations représentant les fonctions syntaxiques. Les relations (binaires pour la plupart ou ternaires) peuvent associer indifféremment des formes individuelles ou des constituants. Notons, qu'EASY ne connait pas la notion de *tête* lexicale (Gendner *et al.*, 2003) (Vilnat *et al.*, 2004).

Dans EASY, il y a 6 types de constituants : (1) nominal, (2) adjectival, (3) prépositionnel, (4) adverbial, (5) verbal et (6) prépositionnel-verbal, le dernier étant utilisé pour les verbes à l'infinitif introduits par une préposition, et 14 types de relations de dépendance : (1) sujet-verbe, (2) auxilliaire-verbe, (3) c-o-d, (4) complément-verbe, (5) modifieur de non, (6) modifieur de verbe, (7) modifieur d'adjectif, (8) modifieur d'adverbe, (9) modifieur de préposition, (10) complémenteur, (11) attribut du sujet/objet, (12) coordination, (13) apposition, (14) juxtaposition. Le choix de ces constituants et de ces relations a été fait à la suite de discussions avec l'ensemble des participants à la campagne. Il a ensuite fait l'objet d'une description plus d'étaillée à la fois pour les participants et pour les annotateurs dans un guide[3], Ils sont également décrits dans (Vilnat *et al.*, 2004). La figure 1 donne un exemple d'annotation d'une phrase issue du corpus littéraire.

Pour comparer les résultats des différents analyseurs, les mesures d'évaluation sont la précision et le rappel (ainsi que la f-mesure qui les combine) sur lesquelles nous avons expérimenté 15 relâchements de contrainte différents (Paroubek *et al.*, 2006), obtenus en combinant les 5 manières présentées dans la table 1 de comparer les empans de textes correspondant soit aux constituants soit aux cibles de relations, avec les 3 façons de considérer les définitions des constituants (ceux de l'hypothèse, ceux de la référence, ou ceux de l'hypothèse lorsqu'ils existent sinon ceux de la

[2]Les transcriptions d'émission radio-télévisées fournies par le projet ESTER de TECHNOLANGUE sur l'évaluation de la transcription de parole automatique n'ont finalement pas été prises en compte dans le calcul des performances en raison d'un problème dans la segementation des énoncés.

[3]Le guide d'annotation est disponible à l'URL www.limsi.fr/Recherche/CORVAL/easy

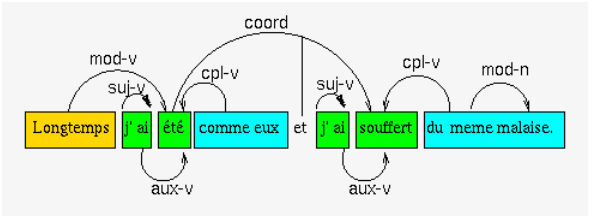


FIG. 1 – Exemple d'annotation d'un énoncé extrait du corpus littéraire (Coppé).

référence). L'évaluation a été menée indépendamment sur les constituants et les relations. Les résultats ont été calculés individuellement pour chaque constituant, chaque relation et chaque type de corpus ainsi que de manière globale.

Fonction	Formule
ÉGALITÉ	$H = R$
FLOU UNITAIRE	$\lvert H \backslash R \rvert \leq 1$
INCLUSION	$H \subset R$
INTERSECTION	$R \cap H \neq \emptyset$
BARYCENTRE	$\frac{2*\lvert R \cap H \rvert}{\lvert R \rvert + \lvert H \rvert} > 0.25$

avec :
H Empan de texte hypothèse
et
R Empan de texte référence,

TAB. 1 – Comparaison des empans correspondant aux constituants et aux cibles des relations.

3 Les résultats de la campagne EASY

Dans tout cette partie qui illustre les résultats des participants, nous ne donnerons pas directementleurs noms, nous y ferons référence par le biais de noms *anonymisés* Pi. Notre but n'est pas de donner un classement de ces participants mais d'indiquer les performances obtenues, ainsi que les écarts observés entre ces performances dans les différents domaines de l'évaluation.

3.1 Les mesures sur les constituants

Pour les constituants c'est le système P10 qui obtient les meilleurs résultats pour les 3 mesures (précision, rappel, f-mesure), tous constituants et tous genres de corpus confondus avec la comparaison barycentre pour les empans de texte des constituants (voir table 1 pour la définition de ces notions). La figure 2 illustre les résultats obtenus par ce participant, avec les différents corpus et les constituants annotés sur le plan horizontal (respectivement axe des x et des y) et la performance calculée en vertical (axe des z). Le graphe de gauche correspond à une vue avant, celui de droite à une vue arrière, comme l'illustrent les petits schémas au-dessus des graphes..

Nous avons utilisé la mesure barycentrique, car c'est celle qui, tout en permettant un certain relâchement des contraintes imposées sur les frontières de constituant (qui sont parfois le résultat d'un choix arbitraire), sans toutefois être aussi laxiste que l'intersection (où il suffit qu'un seul mot soit partagé).

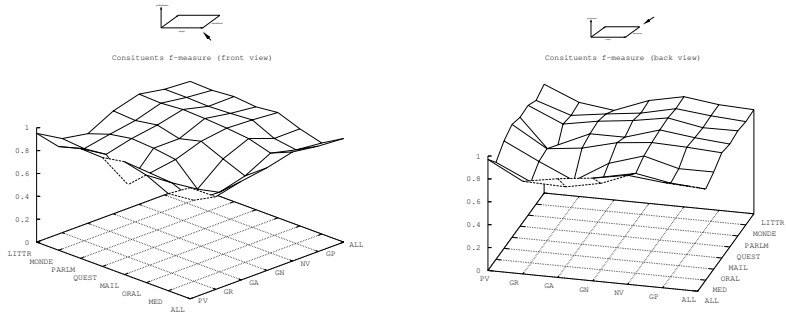


FIG. 2 – Vue avant et arrière sur les performances en f-mesure de P10 pour les constituants.

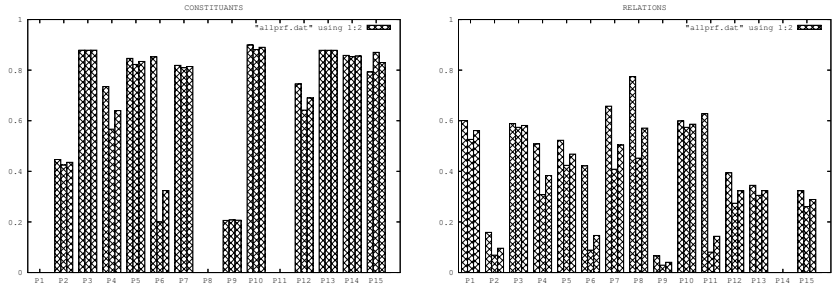


FIG. 3 – Résultats des 15 analyseurs pour les constituants (à gauche) et les relations (à droite) en précision/rappel/f-mesure, tous corpus et toutes annotations confondus

La table 2 donne les résultats de tous les analyseurs par type de corpus pour tous les constituants, en précision et f-mesure, pour distinguer les analyseurs visant à la correction de ceux visant à l'exhaustivité. Comme nous pouvions nous y attendre, les mesures de performance sur les constituants s'apparentent fortement aux types de résultat que l'on obtient avec un simple étiquetage morpho-syntaxique, les problémes étant assez similaires. Le profil des résultats est assez plat et dépend peu du type d'annotation ou du type de corpus traité, au contraire de ce qui se passe pour les relations comme nous le verrons plus loin.

La figure 3 illustre les résultats des différents analyseurs en combinant tous les corpus et toutes les annotations, à la fois en précision, rappel et f-mesure. Sur la figure de gauche, on peut observer 12 colonnes, car trois participants n'ont pas fourni de résultats pour les annotations en constituants mais uniquement l'annotation des relations de dépendance. De même sur la figure de droite, on voit que l'un des participants n'a pas fourni d'annotation en relations de dépendance.

	lemonde	littéraire	médical	oral_delic	parlement	questions	web
P1	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0
P2	p=0.717 f=0.690	p=0.329 f=0.320	p=0.332 f=0.312	p=0.612 f=0.591	p=0.702 f=0.644	p=0.395 f=0.373	p=0.719 f=0.679
P3	p=0.920 f=0.926	p=0.901 f=0.912	p=0.907 f=0.913	p=0.752 f=0.760	p=0.923 f=0.930	p=0.931 f=0.935	p=0 f=0
P4	p=0.813 f=0.660	p=0.802 f=0.770	p=0.459 f=0.436	p=0.787 f=0.717	p=0.808 f=0.653	p=0.877 f=0.856	p=0.841 f=0.696
P5	p=0.883 f=0.878	p=0.847 f=0.824	p=0.882 f=0.873	p=0.714 f=0.713	p=0.876 f=0.868	p=0.901 f=0.894	p=0.877 f=0.880
P6	p=0.837 f=0.782	p=0 f=0	p=0 f=0	p=0 f=0	p=0.849 f=0.803	p=0 f=0	p=0.903 f=0.893
P7	p=0.832 f=0.832	p=0.838 f=0.845	p=0.825 f=0.805	p=0.784 f=0.743	p=0.833 f=0.831	p=0.826 f=0.822	p=0.739 f=0.734
P8	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0
P9	p=0.141 f=0.137	p=0.145 f=0.152	p=0.191 f=0.183	p=0.336 f=0.334	p=0.175 f=0.159	p=0.305 f=0.301	p=0.856 f=0.866
P10	p=0.904 f=0.904	p=0.910 f=0.909	p=0.909 f=0.902	p=0.849 f=0.794	p=0.921 f=0.917	p=0.913 f=0.902	p=0.924 f=0.922
P11	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0
P12	p=0.737 f=0.685	p=0.714 f=0.681	p=0.806 f=0.733	p=0.605 f=0.562	p=0.712 f=0.649	p=0.832 f=0.767	p=0.801 f=0.749
P13	p=0.888 f=0.884	p=0.901 f=0.910	p=0.903 f=0.892	p=0.803 f=0.763	p=0.907 f=0.909	p=0.910 f=0.903	p=0.913 f=0.911
P14	p=0.855 f=0.855	p=0.887 f=0.895	p=0.879 f=0.869	p=0.775 f=0.731	p=0.867 f=0.867	p=0.873 f=0.866	p=0.879 f=0.875
P15	p=0.802 f=0.836	p=0.795 f=0.839	p=0.835 f=0.870	p=0.770 f=0.747	p=0.835 f=0.868	p=0.860 f=0.878	p=0.808 f=0.843

TAB. 2 – Mesures en précision (p) et f-mesure (f) par type de corpus pour tous les constituants

3.2 Les mesures sur les relations

Pour les relations, c'est le système P8 qui obtient la meilleure précision, le système P3 qui obtient le meilleur rappel et le système P10 qui obtient la meilleure f-mesure toutes relations et tous genres de corpus confondus en tenant compte des constituants de l'hypothèse lorsqu'ils existent sinon de ceux de la référence et avec la comparaison barycentre pour les empans de texte des constituants (voir table 1). On voit dans le figure 4 les graphes respectifs de ces trois participants, avec les mêmes conventions que dans la figure 2 .

Le tableau 3 présente les résultats de tous les analyseurs en précision et f-mesure, pour toutes les relations, par type de corpus.

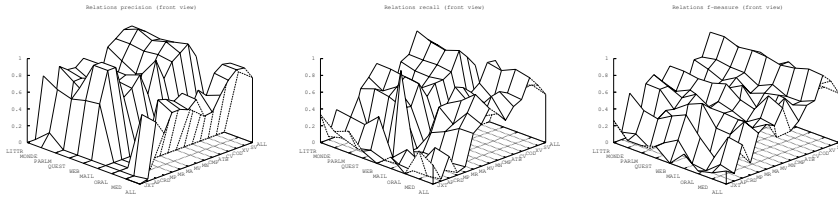


FIG. 4 – Vues avant sur les performances toutes relations et tous genres de corpus confondus pour les meilleures performances en précision (P8), rappel (P3) et f-mesure (P10)

	lemonde	littéraire	médical	oral_delic	parlement	questions	web
P1	p=0.571 f=0.543	p=0.611 f=0.576	p=0.599 f=0.561	p=0.608 f=0.544	p=0.579 f=0.546	p=0.683 f=0.648	p=0.594 f=0.549
P2	p=0.319 f=0.173	p=0.083 f=0.054	p=0.068 f=0.046	p=0.333 f=0.144	p=0.29 f=0.163	p=0.158 f=0.089	p=0.418 f=0.226
P3	p=0.628 f=0.616	p=0.577 f=0.596	p=0.641 f=0.634	p=0.555 f=0.513	p=0.593 f=0.590	p=0.662 f=0.635	p=0 f=0
P4	p=0.583 f=0.409	p=0.529 f=0.429	p=0.277 f=0.231	p=0.563 f=0.459	p=0.551 f=0.400	p=0.669 f=0.607	p=0.554 f=0.415
P5	p=0.562 f=0.508	p=0.507 f=0.456	p=0.564 f=0.524	p=0.514 f=0.425	p=0.529 f=0.472	p=0.447 f=0.412	p=0.553 f=0.489
P6	p=0.419 f=0.377	p=0 f=0	p=0 f=0	p=0 f=0	p=0.410 f=0.372	p=0 f=0	p=0.463 f=0.433
P7	p=0.663 f=0.521	p=0.681 f=0.524	p=0.652 f=0.527	p=0.633 f=0.434	p=0.644 f=0.498	p=0.665 f=0.521	p=0.608 f=0.472
P8	p=0.762 f=0.656	p=0.797 f=0.651	p=0.790 f=0.699	p=0 f=0	p=0.746 f=0.644	p=0.771 f=0.696	p=0.795 f=0.686
P9	p=0.004 f=0.003	p=0.023 f=0.015	p=0.042 f=0.026	p=0.257 f=0.128	p=0.003 f=0.002	p=0.110 f=0.065	p=0.688 f=0.416
P10	p=0.610 f=0.599	p=0.640 f=0.624	p=0.605 f=0.597	p=0.522 f=0.502	p=0.582 f=0.568	p=0.635 f=0.622	p=0.595 f=0.573
P11	p=0.604 f=0.131	p=0.640 f=0.160	p=0.622 f=0.169	p=0.646 f=0.175	p=0.597 f=0.137	p=0.605 f=0.161	p=0.670 f=0.111
P12	p=0.406 f=0.338	p=0.389 f=0.320	p=0.433 f=0.375	p=0.337 f=0.258	p=0.365 f=0.289	p=0.483 f=0.402	p=0.406 f=0.337
P13	p=0.355 f=0.338	p=0.429 f=0.404	p=0.359 f=0.343	p=0 f=0	p=0.337 f=0.321	p=0.354 f=0.330	p=0.268 f=0.255
P14	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0
P15	p=0.336 f=0.312	p=0.381 f=0.340	p=0.326 f=0.302	p=0 f=0	p=0.335 f=0.311	p=0.358 f=0.319	p=0.337 f=0.329

TAB. 3 – Mesures en précision (p) et f-mesure (f) par type de corpus pour toutes les relations

4 Les leçons à tirer

Tout d'abord, rappelons que ce n'est pas parce qu'un système a une valeur de performance 0 pour un sous-corpus ou une relation particulière qu'il a de mauvaises performances, il peut

genre	énoncés nb total	mots nb total	relations nb total	énoncés erronés/testés	relations erronées/testées
WEB	77	2104	113	3/7 = 43%	4/77= 03%
LE MONDE	380	10081	5072	12/39= 30%	22/519= 04%
PARLEMENT	276	7551	3884	14/28= 50%	57/366= 15%
LITTÉRATURE	892	24358	12725	36/93= 38%	92/1196= 07%
EMAILS	852	9243	3960	21/75= 28%	30/421= 07%
MÉDICAL	554	11799	5595	16/54= 29%	28/518= 05%
ORAL_DELIC	505	8117	4591	10/50= 20%	14/462= 03%
QUESTIONS	203	4116	2165	9/20= 45%	20/217= 09%

TAB. 4 – Nombres d'énoncés et de mots par genre de sous-corpus dans la référence.

s'agir d'un choix délibéré de son concepteur de ne pas traiter un phénomène particulier ou de ne retourner qu'une sorte d'annotation, par exemple seulement les relations. Ensuite, de mauvaises performances peuvent provenir de problèmes d'alignement entre les données du participant et celles de références et non d'un mauvais analyseur. Rappelons que dans EASY, contraitement à ce qui avait été fait dans GRACE (Adda *et al.*, 1999) ou dans (Roark *et al.*, 2006), il n'y a pas de procédure de réalignement automatique des données du participant, celui-ci doit respecter la segmentation en mots et en phrases des données qu'il traite.

Concernant les résultats, nous constatons, comme cela était à prévoir, une plus grande variabilité et de moins bonnes performances pour les relations que pour les constituants. Bien entendu, ces résultats ne sont qu'un point de vue ponctuel et sont à relativiser (comme dans toute évaluation quantitative) en fonction des facteurs décrits ci-après. Tout d'abord, la qualité des annotations de référence : nous avons réalisé une première estimation du taux d'erreur d'annotation sur les relations, par type de corpus en demandant à un expert d'examiner à la main un échantillon représentant environ un dixième de chaque corpus annoté. Les résultats sont donnés dans la table 4. Un énoncé est considéré comme erroné s'il contient au moins une erreur d'annotation en relation.

Pour les sous-corpus ayant un taux d'erreur en relation supérieur à 6%, nous avons effectué des corrections systématiques des erreurs les plus fréquentes avant de lancer les calculs de performance [4]. L'estimation du taux d'erreur d'annotation pour tous les sous-corpus permettra de déterminer des classes de performance parmi les différents systèmes sans prendre en compte des différences de performance inférieures aux taux d'erreur estimé.

Le second point dont il faut tenir compte concerne les erreurs de segmentation en mots/phrases encore présentes dans la référence et qui nous ont conduit en particulier à abandonner le traitement du corpus oral provenant de la campagne ESTER. Ces erreurs (auxquelles parfois s'ajoutent les erreurs de format des données des participants) sont à notre avis le résultat de divers facteurs : l'absence de tests *à blanc* du protocole (par manque de temps) et le fait d'avoir imposé une segmentation en mots et phrases de la référence, qui se heurte au problème de déterminer une définition acceptable par tous.

Dans le projet PASSAGE, qui regroupe certains des participants d'EASY, nous annoterons un

[4] Pour le moment nous n'avons pas estimé le taux d'erreur d'annotation pour les sous-corpus web et emails, ni effectué une nouvelle estimation pour les sous-corpus dont les erreurs les plus fréquentes ont été corrigées.

grand corpus en combinant automatiquement des analyseurs syntaxiques. Pour les deux campagnes d'évaluation prévues, nous envisageons de recourir à des procédure d'alignement automatique à partir du texte comme dans GRACE (Adda *et al.*, 1999) ou (Roark *et al.*, 2006). Les participants pourront ainsi conserver leurs propres algorithmes de segmentation en mots et phrases. La phrase dans les données de référence sera déterminée à partir des annotations elles-mêmes, une phrase étant constituée par l'empan de texte sur lequel un arbre syntaxique se projette, comme cela a déjà été fait dans EASY pour le sous-corpus ORAL-DELIC.

Bien entendu, le formalisme d'annotation EASY s'il semble suffisament abouti pour les relations les plus fréquentes comme la relation sujet-verbe, nécessite d'être approfondi pour les autres ; ce qui sera fait dans le cadre du projet PASSAGE, où cette fois nous considèrerons des constitutants admettant plusieurs niveaux de récursivité.

5 Conclusion

EASY a permis de poser les bases d'un protocole d'évaluation des analyseurs syntaxiques du français en mode boîte noire avec des mesures quantitatives objectives. Il a surtout été l'occasion de former un groupe autour du problème de l'évaluation comparative des technologies d'analyse syntaxique et d'acquérir une première expérience dans le cadre d'une campagne d'envergure qui déjà trouve des prolongements dans le projet PASSAGE. Concernant les mesures de performances proprement dites, l'image ponctuelle qu'elles donnent des performances des analyseurs syntaxiques à un instant particulier, nous montre qu'il reste encore un fort potentiel de développement dans la combinaison des approches pour l'annotation de relations syntaxiques, car ce sont 3 systèmes différents qui obtiennent chacun les meilleurs résultats pour la précision, le rappel et la f-mesure. Ce qui laisse à penser que ces systèmes ont des caractéristiques complémentaires, il reste encore à les identifier et à trouver le moyen de les combiner harmonieusement.

Références

ABEILLÉ A. (1991). Analyseurs syntaxiques du français. *Bulletin Semestriel de l'Association pour le Traitement Automatique des Langues*, **32**, 107–120.

ABEILLÉ A., CLÉMENT L. & KINYON A. (2000). Building a treebank for french. In *Proceedings of the 2nd International Conference on Language Ressources and Evaluation (LREC)*, p. 1251–1254, Athen, Greece.

ADDA G., MARIANI J., PAROUBEK P., RAJMAN M. & LECOMTE J. (1999). L'action grace d'évaluation de l'assignation des parties du discours pour le français. *Langues*, **2**(2), 119–129.

BLACHE P. & MORIN J. (2003). Une grille d'évaluation pour les analyseurs syntaxiques. In *Acte de l'atelier sur l'Evaluation des Analyseurs Syntaxiques dans les actes de la 10e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, Batz-sur-Mer.

BLACK E., ABNEY S., FLICKENGER D., GDANIEC C., GRISHMAN R., HARISON P., , HINDLE D., INGRIA R., JELINECK F., KLAVAN J., LIBERMAN M., MARCUS M., ROUKOS S., SANTORINI B. & STRZALKOZSKIJL (1991). A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, p. 306–311, Pacific Grove, California : Morgan Kaufman.

BRANT S., DIPPER S., HANSEN S., LEZIUS W. & SIMTH G. (2002). The tiger treebank. In *Proceedings of the 1st Workshop on Treebank and Linguistics Thories (TLT)*, Sozopol, Bulgaria.

CARROLL J., LIN D., PRESCHER D. & USZKOREIT H. (2002). Proceedings of the workshop beyond parseval - toward improved evaluation measures for parsing systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

GAIZAUSKAS R., HEPPLE M. & HUYCK C. (1998). Modifying existing annotated corpora for general comparative evaluation of parsing. In *Proceedings fo the Workshop on Evaluation of Parsing Systems in the Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

GENDNER V., ILLOUZ G., JARDINO M., MONCEAUX L., PAROUBEK P., ROBBA I. & VILNAT A. (2003). Peas the first instanciation of a comparative framework for evaluating parsers of french. In *Proceedings of the 10th Conference of the European Chapter fo the Association for C omputational Linguistics*, p. 95–98, Budapest, Hungarie. Companion Volume.

MARCUS M., SANTORINI B. & MARCINKIEWCIZ M. (1993). Building a large annotated corpus of english : The penn treebank. *Computational Linguistics*, **19**, 313–330.

MUSILLO G. & SIMA'AN K. (2002). Toward comparing parsers from different linguistic frameworks - an information theoretic approach. In *Proceedings of the Workshop Beyon Parseval - Toward improved evaluation measures for parsing systems at the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

OEPEN S., NETTER K. & KLEIN J. (1996). Test suites for natural language processing. In *CSLI Lecture Notes*. Center for the Study of Language and Information.

PAROUBEK P., POUILLOT L.-G., ROBBA I. & VILNAT A. (2005). Easy : Campagne d'évaluation des analyseurs syntaxiques. In *Proceedings of the 12e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 3–12, Dourdan, France.

PAROUBEK P., ROBBA I., VILNAT A. & AYACHE C. (2006). Data, annotations and measures in EASY - the evaluation campaign for parsers of French. In ELRA, Ed., *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, p. 315–320, Genoa, Italy : ELRA.

ROARK B. (2002). Evaluating parser accuracy using edit distance. In *Proceedings of the Workshop Beyon Parseval - Toward improved evaluation measures for parsing systems at the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

ROARK B., HARPER M., CHARNIAK E., DORR B., JOHNSON M., KAHN J., LIN Y., OSTENDORF M., HALE J., KRANYANSKAYA A., LEASE M., SHAFRAN I., SNOVER M., STEWARD R. & YUNG L. (2006). Sparseval : Evaluation metrics for parsing speech. In *Proceedings of the 5th International Conference on Language Ressources and Evaluation (LREC)*, Genoa, Italy.

SRINIVAS B., DORAN C., HOCKEY B. & JOSHI K. (1996). An approach to robust partial parsing and evaluation metrics. In *Proceedings of the Workshop on Robust Parsing*, Pragues : ESSLI.

VILNAT A., PAROUBEK P., MONCEAUX L., ROBBA I., GENDNER V., ILLOUZ G. & JARDINO M. (2004). The ongoing evaluation campaign of syntactic parsing of french : Easy. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, p. 2023–2026, Lisboa, Portugal.

Modèles statistiques enrichis par la syntaxe pour la traduction automatique

Holger Schwenk, Daniel Déchelotte
Hélène Bonneau-Maynard, Alexandre Allauzen
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex
{schwenk,dechelot,hbm,allauzen}@limsi.fr

Résumé. La traduction automatique statistique par séquences de mots est une voie prometteuse. Nous présentons dans cet article deux évolutions complémentaires. La première permet une modélisation de la langue cible dans un espace continu. La seconde intègre des catégories morpho-syntaxiques aux unités manipulées par le modèle de traduction. Ces deux approches sont évaluées sur la tâche Tc-Star. Les résultats les plus intéressants sont obtenus par la combinaison de ces deux méthodes.

Abstract. Statistical phrase-based translation models are very efficient. In this paper, we present two complementary methods. The first one consists in a a statistical language model that is based on a continuous representation of the words in the vocabulary. By these means we expect to take better advantage of the limited amount of training data. In the second method, morpho-syntactic information is incorporated into the translation model in order to obtain lexical disambiguation. Both approaches are evaluated on the Tc-Star task. Most promising results are obtained by combining both methods.

Mots-clés : traduction automatique, approche statistique, modélisation linguistique dans un espace continu, analyse morpho-syntaxique, désambiguïsation lexicale.

Keywords: statistical machine translation, continuous space language model, POS tagging, lexical disambiguation.

1 Introduction

La traduction automatique est un thème de recherche depuis plusieurs décennies et différentes approches ont été proposées, telles que la traduction par règles, la traduction à base d'exemples ou la traduction statistique. Les travaux récents en traduction statistique confirment que les modèles fondés sur des séquences de mots (Och *et al.*, 1999; Koehn *et al.*, 2003) obtiennent des performances significativement meilleures que ceux fondés sur des mots (Brown *et al.*, 1993). En utilisant des séquences de mots, les systèmes de traduction parviennent à préserver certaines contraintes locales sur l'ordre des mots. L'entraînement d'un tel modèle nécessite l'alignement d'un corpus parallèle. Les régularités du langage naturel comme celles de la syntaxe, ou, encore à un niveau supérieur, celles de la sémantique sont ainsi, en principe, implicitement capturées par les modèles.

Depuis les débuts de l'approche statistique en traduction automatique, les efforts de modélisation se sont principalement concentrés sur les modèles de traduction et d'alignement, comme en témoignent les nombreuses publications sur ces sujets. Dans cet article, nous explorons deux pistes complémentaires pour l'amélioration des modèles de traduction statistique : d'une part, l'exploration d'une modélisation statistique du langage dans un espace continu, et d'autre part l'intégration d'informations syntaxiques dans le modèle de traduction.

Traditionnellement, les systèmes de traduction statistiques utilisent des modèles de langage trigramme à repli. Dans ces modèles classiques, les mots sont représentés par un indice dans un espace discret, le vocabulaire. Ceci ne permet pas de faire de véritables interpolations des probabilités d'un n-gramme non observé puisqu'un changement dans l'espace des mots peut entraîner un changement arbitraire de la probabilité. Nous proposons ici d'appréhender *dans un domaine continu* le problème de l'estimation d'un modèle linguistique. L'idée consiste à projeter les indices des mots dans une représentation continue (un espace vectoriel) et d'estimer les probabilités dans cet espace (Bengio *et al.*, 2003). Actuellement, un réseau de neurones multi-couches complètement connecté est utilisé pour apprendre conjointement la projection des mots sur un espace continu et l'estimation des probabilités n-grammes.

La lecture humaine des sorties d'un système statistique de traduction, même basé sur des séquences de mots, nécessite parfois un difficile exercice de réordonnancement et de restructuration syntaxique pour restituer le sens de l'énoncé d'origine. La modélisation du langage comme une source markovienne (modèle de langage n-gramme), avec comme unité le mot ou la séquence de mots, ne permet pas de prendre en compte les contraintes syntaxiques ou les dépendances à long terme entre les mots. Il apparaît donc nécessaire d'utiliser des méthodes dans lesquelles les propriétés structurelles des langues sont explicitement représentées. Plusieurs tentatives sur l'utilisation d'informations morpho-syntaxiques dans la traduction statistique ont déjà été menées. (Och *et al.*, 2004) ont exploré de nombreuses fonctions caractéristiques, dont certaines d'ordre syntaxique. La réévaluation des n meilleures hypothèses avec des étiquettes morpho-syntaxiques a également été étudiée par (Hasan *et al.*, 2006). Dans (Kirchhoff & Yang, 2005), un modèle de langage factorisé quadrigramme utilisant des informations syntaxiques n'a pas montré des performances meilleures qu'un modèle n-gramme de mots. Les modèles de langage fondés sur la syntaxe ont enfin été explorés par (Charniak *et al.*, 2003). Tous ces travaux ont en commun d'utiliser des séquences de mots comme unités du système de traduction et de n'introduire les catégories morpho-syntaxiques que dans une seconde passe de traitement.

Dans ce travail, nous proposons d'intégrer les informations syntaxiques *dans* le modèle de traduction lui-même. De plus, nous proposons de combiner cette approche avec les méthodes classiques de réévaluation de listes de n meilleures hypothèses. À notre connaissance, cette approche n'a pas été évaluée sur une large tâche (elle a été appliquée par (Hwang *et al.*, 2007) à la tâche BTEC (Basic Travel Expression Corpus) beaucoup plus réduite). Nous présentons ici des résultats sur la tâche Tc-Star (traduction des transcriptions des sessions plénières du Parlement européen).

Cet article est organisé comme suit. Dans la section suivante, nous présentons d'abord la structure du système de traduction automatique et ses différentes extensions. Les résultats expérimentaux sont résumés et discutés dans la section 3. La dernière section conclue cet article et suggère des extensions et travaux futurs.

2 Description du système

L'objectif d'un système de traduction automatique est de proposer pour une phrase $\mathbf{f}$ en langue « source » sa traduction en une phrase $\mathbf{e}$ dans la langue « cible ». L'approche statistique consiste à choisir, parmi les phrases possibles, la plus probable. Le problème se décompose de la manière suivante :

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{e}),$$

où la probabilité $\Pr(\mathbf{f}|\mathbf{e})$ est estimée par le modèle de traduction et $\Pr(\mathbf{e})$ par le modèle de langage de la langue cible. Cette équation résume l'approche *source/canal* historique (Brown *et al.*, 1993) qui considère le mot comme unité et la phrase comme une séquence de mots. Le modèle de traduction peut être estimé automatiquement à partir de textes parallèles alignés au niveau de la phrase. Ce calcul est effectué par le logiciel libre GIZA++.

Ces dernières années, les travaux en traduction statistique ont étendu avec succès l'unité qu'était le mot à la séquence de mots (Och *et al.*, 1999; Koehn *et al.*, 2003). Cette nouvelle unité se définit alors comme un groupe de mots successifs $\tilde{\mathbf{f}}$ de la langue source. Sa traduction est également une séquence de mots $\tilde{\mathbf{e}}$ dans la phrase cible. Les séquences de mots peuvent être extraites automatiquement à partir de données bilingues alignées au niveau du mot dans les deux sens. L'utilisation du principe du maximum d'entropie permet de décomposer le problème de la manière suivante (Och & Ney, 2002) :

$$\mathbf{e}^* = \arg\max p(\mathbf{e}|\mathbf{f}) = \arg\max_{e} \{\exp \sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f})\} \quad (1)$$

où chaque fonction h_i quantifie l'adéquation des phrases $\mathbf{f}$ et $\mathbf{e}$[1]. Les coefficients λ_i pondèrent l'importance relative de ces fonctions.

2.1 Décodeur Moses

Moses[2] est un système de traduction automatique à base de séquences de mots à l'état de l'art. Il est distribué librement avec les scripts nécessaires à l'entraînement d'un système de traduction complet, ainsi qu'une mise en œuvre efficace d'un algorithme de recherche de type *recherche en faisceau* pour produire les traductions. Le décodeur Moses peut également générer une liste des n hypothèses envisagées les plus probables. Cette liste des n meilleures hypothèses contient en général plusieurs fois la même phrase, avec des probabilités différentes, puisque plusieurs segmentations de la phrase source en séquences de mots peuvent aboutir à une même phrase cible. Comme effectué dans les expériences ci-dessous, il est possible de contraindre le décodeur pour que cette liste contienne n hypothèses distinctes.

Dans sa version standard, Moses utilise huit fonctions caractéristiques modélisant le processus de traduction. Ces fonctions permettent d'intégrer à la recherche de la phrase cible les contraintes suivantes : les probabilités de traduction des séquences de mots dans les

[1] Cette « adéquation » est à prendre au sens large, puisqu'un système de traduction inclut toujours un modèle de langage cible $h_i(\mathbf{e}, \mathbf{f}) = p(\mathbf{e})$.

[2] `http://www.statmt.org/moses/`

deux sens, les probabilités de traduction des mots dans les deux sens, une mesure de distorsion, deux pénalités d'insertion de mots et de séquences de mots, et la probabilité calculée par le modèle de langage de la langue cible.

L'approche couramment employée pour optimiser les poids λ_i des fonctions caractéristiques est la maximisation sur un corpus de développement de la mesure BLEU (Papineni *et al.*, 2002). Pour cela, l'outil d'optimisation numérique *Condor* (Berghen & Bersini, 2005) est intégré à l'algorithme itératif suivant :

1. Partant d'un jeu de poids initial, les listes des $n = 1000$ meilleures hypothèses sont générées avec Moses (une liste par phrase source).
2. Ces listes sont réévaluées en utilisant le jeu de poids courant.
3. Les meilleures hypothèses sont extraites et évaluées.
4. À partir du score BLEU aisni calculé, *Condor* calcule un nouveau jeu de poids (l'algorithme retourne alors à l'étape 2), sauf si un maximum local est détecté ce qui met fin à l'algorithme.

Le jeu de poids solution est en général trouvé après une centaine d'itérations. Remarquons que les listes des 1000 meilleures hypothèses sont générées une seule fois lors de l'initialisation et que les itérations réévaluent les listes des 1000 meilleures hypothèses en fonction des poids proposés par *Condor*.

2.2 Désambiguïsation lexicale par catégories syntaxiques

D'une langue à l'autre, les structures et les propriétés syntaxiques diffèrent, par exemple l'espagnol est une langue fortement fléchie alors que l'anglais l'est peu. Or ces structures syntaxiques induisent des ambiguïtés lexicales qui ne sont pas explicitement prises en compte par la modélisation statistique du processus de traduction décrit dans la section ci-dessus.

Il est toujours possible d'utiliser des modèles de langage n-grammes de catégories morpho-syntaxiques pour réévaluer les listes des n meilleures hypothèses de mots générées par un système de traduction. Ce processus nécessite alors d'étiqueter les hypothèses contenues dans les listes. Cependant, les étiqueteurs morpho-syntaxiques ont été appris sur des énoncés correctement formés, ce qui n'est pas toujours le cas des hypothèses provenant d'un système de traduction automatique. Cette étape peut ainsi être une source d'erreurs qui limite les performances de la réévaluation. Nous proposons donc d'intégrer les catégories morpho-syntaxiques *au cœur* du modèle de traduction, ce qui permet d'éviter cet écueil. L'étiqueteur est alors utilisé sur des énoncés syntaxiquement corrects (en tout cas, des énoncés réellement produits), ici sur les corpus parallèles. Par ailleurs, utiliser lors de l'apprentissage des corpus étiquetés morpho-syntaxiquement dans les deux langues permet de prendre en compte les spécificités syntaxiques des deux langues et leur interaction, alors que dans le cas de la réévaluation des listes de meilleures hypothèses, seules les spécificités de la langue cible interviennent.

Nous proposons d'utiliser dans le modèle de traduction des **unités enrichies** constituées des formes de surface des mots, auxquelles sont agglutinées leurs catégories morpho-syntaxiques respectives. Cette méthode permet une désambiguïsation des mots tenant

compte de leurs rôles et de leurs contextes grammaticaux. Un exemple d'énoncé, avec les unités enrichies, est donné à la Figure 1 en anglais et en espagnol.

Anglais : I_{PP} declare_{VVP} resumed_{VVD} the_{DT} session_{NN} of_{IN} the_{DT} European_{NP} Parliament_{NP}

Espagnol : declaro_{VLfin} reanudado_{VLadj} el_{ART} período_{NC} de_{PREP} sesiones_{NC} del_{PDEL} Parlamento_{NC} Europeo_{ADJ}

FIG. 1 – Exemple d'un texte parallèle composé d'unités enrichies utilisé pour entraîner le modèle de traduction.

Lorsque les modèles de traduction et de langage sont fondés sur les unités enrichies, le système de traduction attend en entrée et produit en sortie des séquences d'unités enrichies. Ainsi les phrases à traduire doivent être préalablement étiquetées. Réciproquement, si une traduction classique est requise en sortie, il est nécessaire de retirer les catégories morpho-syntaxiques de l'hypothèse proposée.

Par ailleurs, il devient possible, sur la base des n meilleures hypothèses enrichies, d'effectuer une réévaluation en utilisant un modèle n-gramme de catégories morpho-syntaxiques, sans avoir à utiliser *a posteriori* un étiqueteur sur ces hypothèses.

Pour les expériences présentées dans cet article, nous avons utilisé *TreeTagger* (Schmid, 1994), un étiqueteur markovien utilisant des arbres de décision pour estimer les probabilités trigramme de transition. Ce logiciel est librement disponible pour les deux langues considérées dans cet article. La version anglaise a été entraînée sur le corpus *PENN treebank*[3], et la version espagnole sur le corpus *CRATER*[4]. Le nombre de catégories est assez restreint : 59 pour l'anglais et 69 pour l'espagnol. Notons que les catégories espagnoles ne contiennent pas de distinction en genre et en nombre.

2.3 Modèle de langage neuronal

L'architecture du modèle de langage neuronal est résumée à la Figure 2. Un réseau de neurones multi-couches complètement connecté est utilisé pour apprendre conjointement la projection des mots dans un espace continu et l'estimation des probabilités n-grammes.

Les entrées du réseau sont les $n-1$ mots précédents du vocabulaire et les sorties sont les probabilités a-posteriori pour *tous* les mots du vocabulaire :

$$P(w_j = i | w_{j-n+1}, ..., w_{j-2}, w_{j-1}) = P(w_j = i | h_j) \qquad \forall i \in [1, N] \tag{2}$$

où N est la taille du vocabulaire et h_j le contexte $w_{j-n+1}, ..., w_{j-1}$. Ces entrées sont projetées sur un espace continu (couche P dans la Figure 2). Les autres couches servent à l'estimation non-linéaire des probabilités. La valeur de la i-ème sortie correspond à la probabilité du n-gramme $P(w_j = i | h_j)$. Le réseau calcule donc directement les probabilités de *tous* les mots du vocabulaire pour le même contexte. L'apprentissage se fait par rétro-propagation du gradient, en utilisant la cross-entropie comme fonction d'erreur.

[3] `http://www.cis.upenn.edu/~treebank`
[4] `http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html`

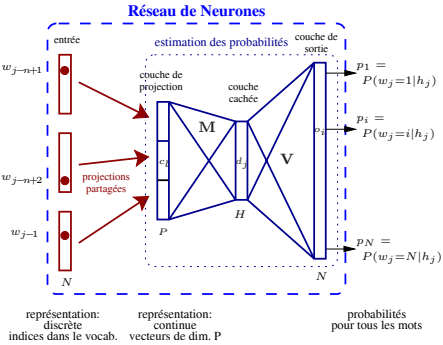


Fig. 2 : Architecture du modèle de langage neuronal. h_j dénomme le contexte $w_{j-n+1}, \ldots, w_{j-1}$. P est la taille d'une projection, et H et N correspondent à la dimension de la couche cachée et de sortie, respectivement.

Dans ce modèle, la complexité est dominée par la taille importante de la couche de sortie. Ainsi, nous proposons de limiter l'estimation des probabilités aux 8 192 mots les plus fréquents, les autres mots étant traités par le modèle à repli standard. Dans nos expériences, environ 90% des requêtes de probabilités sont traitées par le réseau de neurones. Il est important de noter que tous les mots du vocabulaire sont considérés à l'entrée du réseau.

Ce modèle de langage a été utilisé avec succès dans un système de reconnaissance de la parole à grand vocabulaire (Schwenk, 2007), et dans un système de traduction statistique pour la tâche Btec avec un nombre très limité de données d'apprentissage (Schwenk *et al.*, 2006). Cet article décrit la première application du modèle de langage neuronal dans un système de traduction statistique avec plusieurs milliers d'exemples d'apprentissage.

3 Résultats expérimentaux

Les expériences décrites dans cet article ont été effectuées dans le cadre des évaluations internationales organisées par le projet européen Tc-Star[5]. L'objectif de ce projet est de motiver, fédérer, et promouvoir les recherches sur la traduction automatique de la parole. La tâche principale de ce projet est la traduction des transcriptions des sessions plénières du Parlement européen (SPPE). La communauté européenne met à disposition les minutes de ces sessions en plusieurs langues, aussi connues sous le nom « Éditions du texte final » (ETF). Ces textes, alignés au niveau des phrases, sont utilisés pour apprendre les modèles statistiques. Nous disposons également d'environ 100 heures d'enregistrement des sessions plénières du Parlement européen. Ces données audio ont été transcrites manuellement et servent principalement au développement des systèmes de reconnaissance de la parole, mais elles sont aussi utilisées pour entraîner les modèles de langage cible dans le système de traduction.

Trois conditions sont considérées dans les évaluations Tc-Star : la traduction des minutes ETF (*texte*), la traduction des transcriptions des données acoustiques (*verbatim*) et la traduction des hypothèses du système de reconnaissance de la parole (*parole*). Dans ce travail, nous ne considérons que la condition *verbatim*, pour la paire de langues espagnol/anglais. Nous donnons des résultats sur les données de développement et de test de

[5] http://www.tc-star.org/

l'évaluation organisée en 2007. Deux traductions de référence sont disponibles pour les deux jeux de test. Plusieurs étapes de normalisation ont été appliquées aux minutes des sessions plénières afin d'approcher la condition *verbatim* ou *parole*, notamment la transformation en mots des nombres. Les modèles de traduction sont estimés sur les données SPPE qui représentent 1,2M de phrases parallèles, soit environ 35M de mots en anglais.

3.1 Apprentissage des modèles de langage

Pour l'apprentissage des modèles de langage, nous avons utilisé la partie monolingue des données parallèles SPPE ainsi que les transcriptions des données acoustiques. Des données extérieures ont également été utilisées pour une estimation plus robuste des modèles : deux corpus de textes provenant des parlements espagnol (49M mots) et britannique (55M mots). Ainsi, pour chaque langue, nous disposons de trois sources de texte donnant lieu à l'estimation de trois modèles indépendants. Ces trois modèles sont *in fine* interpolés linéairement pour créer un modèle de la langue cible. Les coefficients d'interpolation sont estimés via l'algorithme E.M. de manière à minimiser la perplexité sur les données de développement. Les coefficients obtenus sont 0,81 pour le modèle SPPE, 0,12 pour le modèle estimé sur les données additionnelles du parlement et 0,07 pour celui utilisant les transcriptions acoustiques.

Tous les modèles de langage n-grammes utilisés, hormis le modèle neuronal, sont des modèles classiques avec repli utilisant le lissage de Kneser-Ney modifié. Le SRI LM-toolkit (Stolcke, 2002) a été utilisé pour leur construction.

Les caractéristiques des données et les perplexités des modèles de langage sont résumées dans le Tableau 1. Les modèles trigrammes interviennent pendant le décodage, alors que les modèles quadrigrammes sont utilisés pour réévaluer les listes de n meilleures hypothèses. Le modèle de langage neuronal obtient une réduction de la perplexité de 15% environ. Il est à noter que les données de développement en anglais, donc la traduction de l'espagnol vers l'anglais, proviennent de deux sources différentes (parlements européen et espagnol). Cette différence explique les perplexités relativement élevées. Les perplexités sur les données du Parlement européen uniquement sont plus basses : 85.0, 77.8 et 64.3 pour le tri-, quadrigramme à repli et le quadrigramme neuronal respectivement.

	Anglais	Espagnol
Textes du Parlement européen	35,3M	36,6M
Textes parlementaires supplémentaires	55,1M	48,9M
Transcriptions acoustiques	1,5M	777k
Vocabulaire	82,6k	132,5k
Perplexité trigramme	134,5	69,7
Quadrigramme à repli	123,4	64,0
Quadrigramme neuronal	102,8	54,6

Tab. 1 – Données d'apprentissage (en nombre de mots) utilisées pour l'estimation des modèles de langage et perplexités obtenues sur les données de développement.

3.2 Résultats sur les données de développement

Nous avons effectué de nombreuses études comparatives sur les données de développement pour évaluer les apports des différentes techniques. Les résultats principaux sont résumés dans le Tableau 2. En ce qui concerne la désambiguïsation lexicale, seul le sens de traduction de l'anglais vers l'espagnol (vers la langue la plus infléchie) a été évalué à ce jour. Pour chaque sens de traduction, le score Bleu du modèle de base avec un trigramme est donné, ainsi qu'après la réévaluation avec un quadrigramme à repli et neuronal.

L'utilisation d'un quadrigramme permet d'augmenter le score Bleu d'environ 0,4 points pour la traduction vers l'anglais et de 0,6 points vers l'espagnol. Nous avons également essayé de réévaluer les n meilleures hypothèses avec des modèles de langage n-grammes de catégories morpho-syntaxiques, mais sans effet sur les performances du système. L'utilisation du modèle de langage neuronal, par ailleurs, produit une amélioration du score BLEU de plus de 0,6 points pour les deux directions.

	Espagnol → anglais			Anglais → espagnol					
	Sans désambiguïsation			Sans désambiguïsation			Avec désambiguïsation		
	base	4-gram	NNLM	base	4-gram	NNLM	base	4-gram	NNLM
BLEU	47,20	47,64	48,26	48,78	49,39	50,15	48,92	49,45	50,30

Tab. 2 – Scores BLEU sur les données de développement. NNLM dénomme le modèle de langage neuronal.

Les gains apportés par la désambiguïsation lexicale par catégories syntaxiques sont relativement faibles lorsqu'on considère les systèmes avec un tri- ou quadrigramme à repli. Là encore, une réévaluation avec des modèles n-grammes de catégories syntaxique n'est pas efficace. Cependant, les résultats sont intéressants lorsqu'on combine la modélisation de langage neuronal et la désambiguïsation lexicale : le score Bleu passe de 49,39 à 50,30. Ceci montre bien l'intérêt de travailler conjointement sur une amélioration des techniques statistiques et sur l'incorporation de connaissances lexicales ou syntaxiques. En effet, la réévaluation des n meilleures hypothèses avec un modèle de langage semble être plus efficace si les mots proposés par le modèle de traduction sont mieux choisis.

3.3 Résultats sur les données de test

Les performances sur les données de test de l'évaluation Tc-Star 2007 sont résumées dans le Tableau 3. Les coefficients λ_i des fonctions caractéristiques sont les mêmes que ceux du système optimisé sur les données de développement. Le système n'a donc pas été adapté sur les données de test. Sept centres de recherche publiques et industriels ont participé à l'évaluation qui s'est déroulée en février 2007. Les scores BLEU varient entre 42.95 et 49.60 (espagnol/anglais) et entre 37.39 et 51.04 (anglais/espagnol). Les performances du système avec désambiguïsation lexicale sont très légèrement au-dessous du système de base, dans le cas de l'utilisation d'un modèle de langage à repli. Cependant la combinaison avec un modèle de langage neuronal donne de bons résultats, sans pour autant pouvoir dépasser le système sans désambiguïsation.

	Espagnol → anglais			Anglais → espagnol					
	Sans désambiguïsation			Sans désambiguïsation			Avec désambiguïsation		
	base	4-gram	NNLM	base	4-gram	NNLM	base	4-gram	NNLM
BLEU	48,42	48,67	49,19	49,19	50,17	51,04	49,13	49,91	51,04

TAB. 3 – Scores BLEU sur les données de test.

4 Conclusion

Nous avons présenté et évalué deux évolutions d'un système de traduction statistique. L'une propose une modélisation linguistique dans un espace continu et la seconde intègre les catégories morpho-syntaxiques des mots dans le modèle de traduction. La combinaison des deux méthodes donne des résultats intéressants. Notre système a obtenu de très bons résultats à l'évaluation TC-STAR organisée début 2007.

Nous étudions aussi l'application des mêmes techniques à la traduction automatique d'autres paires de langues, notamment la traduction entre l'anglais et le français. Pour cela le corpus Europarl est utilisé (Koehn, 2006). Nous sommes en train de produire une deuxième référence de traduction qui sera librement disponible pour d'autres laboratoires de recherche intéressés dans la traduction automatique du français[6].

Plusieurs extensions du système décrit dans cet article sont actuellement à l'étude. Nous travaillons sur une meilleure incorporation des connaissances linguistiques, notamment sur l'utilisation d'étiqueteurs prenant en compte le genre et le nombre, voire le sens des mots, afin d'améliorer la désambiguïsation dans le modèle de traduction. Un logiciel de visualisation des erreurs de traduction est en cours de développement afin de permettre une analyse qualitative des erreurs pour affiner le choix des étiquettes, notamment pour le français. En ce qui concerne l'amélioration des techniques statistiques, nous sommes très intéressés par une représentation factorisée des mots, incluant notamment des informations morpho-syntaxiques et linguistiques, aussi bien pour le modèle de traduction que pour le modèle de la langue cible.

Remerciements

Ces recherches ont été partiellement financées par le projet européen TC-STAR et par le projet ANR Instar, JCJC06_143038.

Références

BENGIO Y., DUCHARME R., VINCENT P. & JAUVIN C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**(2), 1137–1155.

BERGHEN F. V. & BERSINI H. (2005). CONDOR, a new parallel, constrained extension of powell's UOBYQA algorithm : Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, **181**, 157–175.

[6]Données disponibles à partir de la page internet `http://instar.limsi.fr`

Brown P., Della Pietra S., Della Pietra V. J. & Mercer R. (1993). The mathematics of statistical machine translation. *Computational Linguistics*, **19**(2), 263–311.

Charniak E., Knight K. & Yamada K. (2003). Syntax-based language models for machine translation. In *MT Summit*.

Hasan S., Bender O. & Ney H. (2006). Reranking translation hypothesis using structural properties. In *EACL Workshop on Learning Structured Information in Natural Language Applications*.

Hwang Y., Finch A. & Sasaki Y. (2007). Improving statistical machine translation using shallow linguistic knowledge. *Computer Speech & Language*, **21**(2), 350–372.

Kirchhoff K. & Yang M. (2005). Improved language modeling for statistical machine translation. In *ACL'05 workshop on Building and Using Parallel Text*, p. 125–128.

Koehn P. (2006). Europarl : A parallel corpus for statistical machine translation. In *MT Summit*.

Koehn P., Och F. J. & Marcu D. (2003). Statistical phrased-based machine translation. In *Joint Conference on Human Language Technology and of the North American Chapter of the Asociation for Computational Lingustics*, p. 127–133.

Och F.-J., Gildea D., Khudanpur S., Sarkar A., Yamada K., Fraser A., Kumar S., Shen L., Smith D., Eng K., Jain V., Jin Z. & Radev D. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of the North American Chapter of the Asociation for Computational Lingustics*, p. 161–168.

Och F. J. & Ney H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*, p. 295–302.

Och F. J., Tillmann C. & Ney H. (1999). Improved alignment models for statistical machine translation. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Copora*, p. 20–28.

Papineni K., Roukos S., Ward T. & Zhu W. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of ACL*, p. 311–318.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schwenk H. (2007). Continuous space language models. *Computer Speech and Language*, **21**, 492–518.

Schwenk H., Costa-jussà M. R. & Fonollosa J. A. R. (2006). Continuous space language models for the IWSLT 2006 task. In *International Workshop on Spoken Language Translation*, p. 166–173.

Stolcke A. (2002). SRILM - an extensible language modeling toolkit. In *International Conference on Speech and Language Processing*, p. II : 901–904.

Traitements phrastiques phonétiques pour la réécriture de phrases dysorthographiées

Laurianne SITBON[1, 2], Patrice BELLOT[1], Philippe BLACHE[2]
[1] Laboratoire d'Informatique d'Avignon - Université d'Avignon
[2] Laboratoire Parole et Langage - Université de Provence
{laurianne.sitbon,patrice.bellot}@univ-avignon.fr,
blache@lpl.univ-aix.fr

Résumé. Cet article décrit une méthode qui combine des hypothèses graphémiques et phonétiques au niveau de la phrase, à l'aide d'une réprésentation en automates à états finis et d'un modèle de langage, pour la réécriture de phrases tapées au clavier par des dysorthographiques. La particularité des écrits dysorthographiés qui empêche les correcteurs orthographiques d'être efficaces pour cette tâche est une segmentation en mots parfois incorrecte. La réécriture diffère de la correction en ce sens que les phrases réécrites ne sont pas à destination de l'utilisateur mais d'un système automatique, tel qu'un moteur de recherche. De ce fait l'évaluation est conduite sur des versions filtrées et lemmatisées des phrases. Le taux d'erreurs mots moyen passe de 51 % à 20 % avec notre méthode, et est de 0 % sur 43 % des phrases testées.

Abstract. This paper introduces a sentence level method combining written correction and phonetic interpretation in order to automatically rewrite sentences typed by dyslexic spellers. The method uses a finite state automata framework and a language model. Dysorthographics refers to incorrect word segmentation which usually causes classical spelling correctors fail. Our approach differs from spelling correction in that we aim to use several rewritings as an expression of the user need in an information retrieval context. Our system is evaluated on questions collected with the help of an orthophonist. The word error rate on lemmatised sentences falls from 51 % to 20 % (falls to 0 % on 43 % of sentences).

Mots-clés : réécriture de phrases, dyslexie, automates, correction orthographique.

Keywords: sentence level rewriting, dyslexia, FSM, spell checking.

1 Introduction

koman sapel le mer de bastya ? Si votre système de conversion graphème-phonème et votre conscience phonologique fonctionnent parfaitement, vous devriez avoir une intuition de l'intention de cette question. Peut-être même pouvez-vous y répondre ? En revanche, les systèmes automatiques pourtant très élaborés de questions réponses en resteront cois, dans le meilleur des cas ils répondront *la Méditerranée*, s'ils s'aident d'un correcteur orthographique. La raison principale est que ces systèmes ne sont pas conçus pour prendre en compte un profil linguistique de l'utilisateur, et dans le cas d'utilisateurs avec un handicap de langage ces systèmes ne

sont généralement pas assez robustes. Les performances des correcteurs orthographiques sont par exemple faibles dans le cas où l'auteur est dysorthographique. Les correcteurs orthographiques grand public supposent une segmentation en mots correcte, ce qui les rend très peu efficaces dans des cas d'écriture dysorthographique, comme le montre une étude menée par (James & Draffan, 2004). Or la dysorthographie est un trouble associé à la dyslexie d'origine phonologique [1], qui provient d'un trouble de la conscience phonologique. Ceci implique généralement une écriture essentiellement phonétique et une segmentation en mots souvent erronée. La conscience phonologique permet de découper une séquence de phonèmes en unités sémantiques dans une phrase (Gillon, 2004). Si les unités sémantiques sont identifiées correctement, le passage à l'orthographe se fait ensuite par des voies de conversion phonème-graphème couplées avec des connaissances sur les exceptions orthographiques reliées au mot.

Avant d'envisager un système de réécriture, la première section situe les besoins par rapport aux systèmes auxquels les réécritures s'adressent ainsi qu'aux utilisateurs à travers la constitution d'un corpus d'évaluation . La seconde section pose les bases d'un système dédié aux dysorthographiques qui s'appuie sur une combinaison d'hypothèses graphémiques et phonétiques au niveau de la phrase, ainsi que sa mise en application. La troisième section propose une évaluation de ce système sur le corpus constitué, en comparant ses performances à celles d'un correcteur orthographique.

2 Enjeux de la réécriture

Dans le cadre de la recherche d'information, on pourra admettre plusieurs hypothèses de réécriture de la requête initiale à l'aide d'un modèle robuste capable de pondérer les différentes hypothèses, ce qui est pratiqué dans le cadre de l'expansion de requêtes. De plus, les systèmes utilisant généralement les lemmes de mots de la requête à la place des formes, une hypothèse contenant des fautes d'accord reste acceptable. Ainsi, il s'agit bien de réécriture en vue d'un traitement automatique et non pas de correction orthographique.

2.1 Constitution d'un corpus d'exemples de questions dysorthographiées

Parmi les différents types de système de recherche d'information, nous nous sommes focalisés sur les systèmes de questions réponses car ils exigent une requête formulée en une phrase cohérente. De plus, les enfants sont les plus touchés par les problèmes de dysorthographie et ils formulent généralement leurs requêtes par des questions.

Le corpus que nous avons recueilli est donc un corpus de questions tapées par des enfants dyslexiques (qui sont également dysorthographiques). Ce corpus a été réalisé lors de séances d'orthophonie de huit enfants (entre 9 ans et demi (CE2) et 13 ans (4e)).

Le choix des questions a été guidé par les contraintes d'évaluation du système de questions réponses (SQR) SQuALIA (Gillard *et al.*, 2006) dont nous disposons ainsi que par le vocabulaire restreint des enfants. Nous avons sélectionné des questions factuelles de la campagne d'évaluation Technolangue EQUER (Ayache *et al.*, 2006) pour lesquelles SQuALIA a fourni une bonne réponse (Gillard *et al.*, 2005), soit environ 200 questions sur les 500 proposées. Parmi ces 200 questions, nous avons sélectionné celles dont tous les mots se trouvent dans le lexique de niveau

[1] Selon les études les plus récentes, la dyslexie touche 3 à 5 % des enfants.

cours préparatoire de Manulex (Lété *et al.*, 2004), qui recense les fréquences des mots de manuels scolaires pour différentes classes d'âge, et donne ainsi un aperçu des mots écrits connus par les enfants. Les 5 questions finalement retenues se trouvent dans le tableau 1.

Qui est le maire de Bastia ?
Quel âge a l'abbé Pierre ?
Quelle est la capitale de Terre Neuve ?
Qui est le frère de la princesse Leia ?
Quelle est la monnaie nationale en Hongrie ?

TAB. 1 – Questions de la campagne EQUER retenues pour constituer le corpus.

Les questions ont été saisies au clavier par les 8 enfants de manière semi-spontannée, c'est à dire qu'elles ne leur ont pas été dictées. Pour chaque question, les quatre étapes suivantes ont été suivies par l'orthophoniste :
– la réponse est dite à l'enfant dans une phrase (*Le maire de Bastia s'appelle X*) ;
– elle demande à l'enfant quelle question il poserait pour obtenir cette réponse (*que me demanderais-tu pour que je te réponde X ?*) ;
– l'enfant tape la question qu'il vient de formuler ;
– l'enfant relit la question qu'il vient de taper et éventuellement corrige ce qu'il veut.
Le corpus ainsi obtenu, bien que de taille réduite (37 phrases), est très représentatif car il permet beaucoup d'observations communes aux huit participants. En premier lieu il apparaît clairement que la plupart des observations faites classiquement sur les manuscrits d'enfants dyslexiques ne sont pas validées sur les écrits typographiées. Cela est du non seulement à une organisation motrice différente pour la production écrite (il ne s'agit pas de former les lettres mais de les repérer sur le clavier, où elles apparaissent en majuscules, il n'est donc plus question de latéralisation), mais également à une plus grande motivation pour la frappe au clavier impliquant une plus grande attention au niveau de la production comme de la relecture. Ainsi, on ne rencontre pas de substitutions de lettres dites "miroirs" (*p, b, d, q* ou *m* et *w*, *n* et *u*). De même on n'observe que deux cas d'inversion de lettres, et aucun cas d'inversion de syllabes.

Les erreurs que l'on rencontre sont essentiellement des erreurs de conversion phonème-graphème au niveau de la phrase. Cela signifie à la fois une écriture phonétique mais pas nécessairement simpliste des mots (ainsi, *monnaie* s'écrit *monné, monais, moner, monnaie, moner, monaie* ou *monai*), et une segmentation en mots erronée (*s'appelle* peut s'écrire *ca ple* ou bien *sapel*, et *l'abbé Pierre* s'écrit *labe pierre, labpier, la Bepierre, labepier, labée pierre, l abepier, l'abée pierre* ou *labpier*). On rencontre également des omissions ou substitutions de lettres dans des cas où les phonèmes ne sont pas assez distincts (comme pour *Bastia* ou *monnaie*). Une autre conséquence de l'écriture phonétique est la substitution de certains mots par des homophones (*mer* remplace *maire*).

Par ailleurs on remarque des motifs d'erreurs constants pour chaque individu et propres à chacun. Par exemple pour un même enfant les pronoms interrogatifs souffrent systématiquement d'un remplacement du *u* par une apostrophe (*q'elle* au lieu de *quel*) ou pour un autre enfant d'un ajout d'apostrophe (*qu'el* au lieu de *quel*). Ces régularités pour un même utilisateur suggèrent la possibilité de définir des modèles individuels d'erreurs modélisant les transitions des orthographes erronées vers les orthographes correctes. Cependant la définition d'un modèle générique pour tous les utilisateurs se révèle impossible, étant donné que comme ces exemples le confirment, il existe autant de dyslexies que de dyslexiques. D'autre part la définition de modèles individuels devrait nécessairement être dynamique car les utilisateurs sont généralement en cours d'apprentissage et les type d'erreurs peuvent évoluer.

3 Un système de réécriture dédié

La réécriture peut se faire à l'aide d'un correcteur orthographique étant donné qu'ils proposent généralement plusieurs alternatives pour chaque mot rencontré hors de leur lexique. Cependant comme le démontre l'étude publiée dans (James & Draffan, 2004) les correcteurs grand public ne répondent généralement pas aux besoins spécifiques des dyslexiques, qui ont tendance à produire un mauvais découpage en mots ainsi qu'à la substitution d'homophones (lesquels homophones se trouvent généralement dans le lexique et ne sont donc pas repérés).

Des modèles pour la correction orthographique dédiée ont été proposés. Ainsi (Loosemore, 1991) propose une modélisation globale des erreurs commises par des dyslexiques, arguant que la dyslexie implique des erreurs aggravées mais pas différentes par rapport à celles produites par des non dyslexiques. De la même manière, (Deorowicz & Ciura, 2005) proposent des réseaux de confusions représentés par des automates, où les alternatives sont issues de modèles de confusion graphiques supposés modéliser les causes d'erreurs. On se rend bien compte avec un corpus tel que celui que nous avons recueilli que ces modèles génériques peuvent rapidement prendre des proportions considérables. (Spooner, 1998), toujours en partant de l'idée qu'une erreur commise par un dyslexique ne se différencie que par son niveau de gravité, propose des modèles spécifiques à chaque utilisateur. Le correcteur qu'il implémente à l'aide de ces modèles obtient des performances comparables à celles des correcteurs orthographiques grand public. Enfin, (Toutanova & Moore, 2002) propose une approche qui combine des modèles de lettres et de phonèmes sur les mots, en se basant sur les approches probabilistes de canal bruité introduites par (Brill & Moore, 2000). L'ensemble de ces systèmes permettent de corriger des mots hors d'un lexique mais ne tiennent pas compte des homophones. (Pedler, 2001) propose une détection de telles erreurs à l'aide de contextes syntaxiques et sémantiques, sur la base d'ensembles de confusion.

Cependant toutes ces applications fonctionnent sur le postulat que les séquences de mots sont correctement identifiables, et que les erreurs sont isolées (pour les systèmes utilisant les informations syntaxiques et sémantiques notamment). Cependant, comme le montre l'analyse de notre corpus, un traitement au niveau de la phrase s'impose. La majorité des erreurs étant de nature graphémique et non phonétique, cela suggère un traitement phonétique au niveau de la phrase entière. Cela lève à la fois le problème de la segmentation en mots et celui des homophones. Les outils de la reconnaissance automatique de la parole offrent des performances intéressantes en se fondant sur des modèles de langage.

3.1 Combinaison d'alternatives graphémiques et phonologiques pour l'interprétation

Une fois oralisées, la plupart des phrases de notre corpus deviennent compréhensibles et interprétables par des êtres humains. A partir de ce constat, nous avons émis l'idée d'un système automatique fonctionnant sur ce principe, en enchaînant une phonétisation de la phrase et une transcription du signal ainsi généré. En pratique le passage par un signal audio n'est pas nécessaire, on peut se contenter d'un passage par une séquence de phonèmes en sortie du phonétiseur et en entrée du système de reconnaissance. Le phonétiseur s'appuie sur un ensemble de règles de conversion graphème-phonèmes ainsi que sur un lexique phonétisé de la langue. Le système de reconnaissance s'appuie sur un modèle de langage ainsi qu'un lexique phonétique pour produire plusieurs hypothèses de transcription, dont le contenu verbal se limite au lexique utilisé.

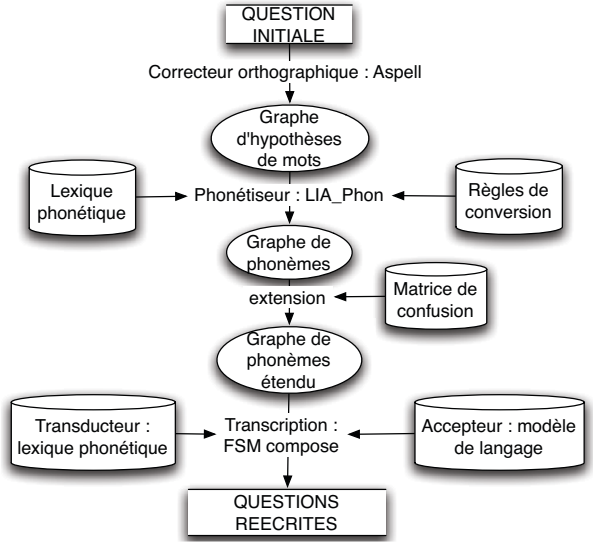


FIG. 1 – Etapes de la réécriture d'une question : état initial et final, représentations intermédiaires, outils pour les transitions entre les représentations, données utilisées par ces outils

En pratique, nous nous sommes rendus compte qu'en se contentant d'une unique séquence de phonèmes correspondant à la phrase, nous perdions trop d'informations. Ainsi les confusions phonétiques (entre les voyelles ouvertes et fermées notamment) n'étaient pas prises en compte. Il faut donc construire un graphe de phonèmes et non pas une séquence de phonèmes. D'autre part, les omissions et les inversions de lettres ne peuvent pas être traitées si l'on s'en tient aux règles de conversion, et générer toutes les possibilités de ce type (en générant des arcs supplémentaires dans le graphe de phonèmes) risquerait d'apporter trop de confusions. Une solution à cela est de générer un graphe d'hypothèses de mots qui représente la séquence écrite, puis de phonétiser toutes les phrases issues de ce graphe de mots afin de générer un graphe de phonèmes plus complet. Les hypothèses de mots peuvent être obtenues à l'aide d'un correcteur orthographique, la plupart se basant sur des distances d'édition. Le graphe de la figure 1 illustre ce processus.

Dans les graphes représentant la phrase aux étapes intermédiaires, les arcs portent les coûts de transition associés aux phonèmes ou aux mots qu'ils portent également, et les noeuds sont les étapes qui séquencent la phrase. Ainsi les différentes phrases hypothèses graphiques ou phonétiques ont un coût associé correspondant à la somme des coûts de transition du chemin emprunté. Le chemin correspondant exactement à ce qui a été écrit doit avoir un coût nul, et plus on s'en écarte plus le coût doit être important.

On attribue un poids Wg aux mots alternatifs H (hypothèses graphémiques) proposés par le correcteur orthographique :

$$Wg(H) = f(d(H, I)), \tag{1}$$

où f est une fonction de normalisation de la distance $d(H, I)$ entre l'hypothèse proposée et le mot initialement écrit. Cette distance peut être fournie par le correcteur, ou calculée *a posteriori* (distance d'édition par exemple).

On attribue un poids Wp aux alternatives phonétiques H (hypothèses phonétiques) obtenues à l'aide d'une matrice de confusion :

$$Wp(H) = g(m(H, I)), \tag{2}$$

où g est une fonction de normalisation d'une distance $m(H, I)$ entre le phonème alternatif et le phonème initial. Cette distance peut faire partie intégrante de la matrice de confusion.

3.2 Mise en application

Le cadre théorique et applicatif proposé par les machines à états finis (FSM) (Mohri *et al.*, 2002) pour la reconnaissance automatique de la parole correspond à nos besoins de représentations intermédiaires en graphes, à travers leur implémentation dans le AT&T FSM Toolkit (Mohri *et al.*, 1997). En effet l'implémentation de modèles de langages dans le formalisme des automates telle que proposée par (Allauzen & Mohri, 2005) avec la bibliothèque *grm* permet de les utiliser pour décoder un automate construit à partir du graphe de phonèmes étendu, à condition de le coupler avec un transducteur permettant de faire correspondre des séquences de phonèmes à des mots écrits. Il s'agit alors de rechercher les meilleurs chemins en fonction des coûts de transition associés dans le graphe composé des hypothèses phonétiques, du modèle de langage et du transducteur déduit du lexique phonétique. Le modèle de langage est appris sur un mois d'articles du journal *Le Monde* disponibles dans le corpus de la campagne EQUER.

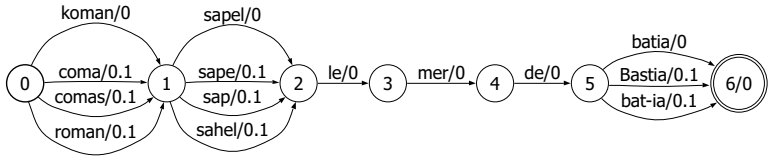


FIG. 2 – Graphe de mots pour la phrase *koman sapel le mer de batia*

Les hypothèses graphémiques permettant de générer le graphe de mots sont obtenues à l'aide du correcteur orthographique libre GNU ASPELL [2]. En mode *badspellers*, il utilise à la fois des distances de Levenshtein (distance d'édition, égale au nombre minimal de caractères qu'il faut supprimer, insérer, ou remplacer pour passer du mot écrit aux mots du lexique) et des distances phonologiques (distance d'édition basée sur les phonèmes) pour proposer des alternatives au mots rencontrés hors de son lexique. Ce correcteur montre de bonnes performances par rapport aux autres outils commerciaux et libres grand public [3]. La figure 2 montre un exemple de graphe de mots ainsi construit. La phonétisation est effectuée à l'aide de l'outil LIA_phon (Bechet, 2001), qui dispose à la fois d'un lexique phonétique de 80 000 mots et d'un système de 1996 règles de conversions ordonnées des plus générales aux plus exceptionnelles. La combinaison

[2] `http://aspell.sourceforge.net`
[3] `http://aspell.net/test/`

de ces deux ressources rend la phonétisation robuste, ce qui est essentiel compte tenu des dégradations orthographiques qui peuvent être rencontrées. La matrice de confusion pour obtenir le graphe de phonèmes étendu contient uniquement les confusions entre les voyelles ouvertes et fermées. Par la suite elle pourra être étendue à l'aide de modèles de confusion phonétiques appris sur un grand corpus. La figure 3 illustre le graphe de phonèmes étendus correspondant au graphe de mots de la figure 2.

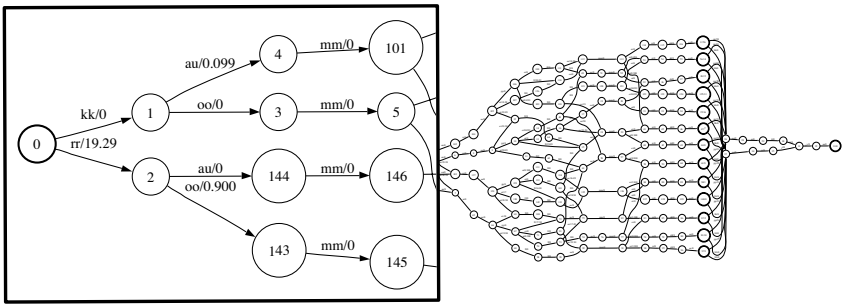


FIG. 3 – Graphe de phonèmes étendu de la phrase *koman sapel le mer de batia*

Phrase	Score
comment s'appelle le maire de bail à	52,2848701
comment s'appelle le maire de bahia	54,6559029
comment s'appelle le maire de bastia	54,8422737

TAB. 2 – Réécritures les plus probables de *koman sapel le mer de batia*

Les fonctions de coût normalisé associées aux alternatives graphémiques ou phonétiques des équations (1) et (2) ont été établies de manière empirique avec les valeurs suivantes :

$$f(d(H,I)) = \begin{cases} 0\ if\ H = I \\ 0.1\ if\ H \neq I \end{cases} \quad (3)$$

$$g(m(H,I)) = \begin{cases} 0\ if\ H = I \\ 0.1\ if\ H \neq I \end{cases} \quad (4)$$

Ainsi la recherche du meilleur chemin dans le graphe final composé du graphe d'hypothèses phonétiques, de l'accepteur du modèle de langage et du transducteur du lexique phonétique prend en compte les coûts de transition issus de chacun des automates, et permet d'affecter des scores à tous les chemins. Le tableau 2 représente les trois hypothèses les plus probables, qui correspondent aux chemins de coût minimal dans le graphe de la figure 3, et les coûts de chacune de ces hypothèses. Dans ce cas, l'hypothèse attendue est la troisième et son score est proche des deux premières. Les phrases improbables ont des scores au delà de 100.

4 Évaluation

Pour l'évaluation, nous avons constitué une référence de la même manière que l'on transcrit manuellement les textes audio pour tester la reconnaissance : nous avons effectué une transcription

manuelle des phrases tapées par les enfants de manière à s'approcher au mieux de leur intention. Il n'y a pas d'ambiguïtés dans les choix de transcription puisque l'on connaît par avance l'objet des questions. La plate forme d'évaluation des outils de reconnaissance de la parole SCTK [4] inclut l'outil SCLITE qui implémente un algorithme de programmation dynamique pour calculer des taux d'erreurs mots dans le meilleur des cas entre une phrase de référence et la phrase correspondante qui peut contenir plusieurs hypothèses (représentées par un graphe de mots), en prenant en compte les insertions, omissions et les substitutions.

Dans le cadre d'une réécriture en entrée d'un système de recherche d'information, il n'est pas nécessaire que tous les mots de la question soient corrects ni qu'ils soient bien accordés. En effet la plupart des systèmes effectuent en premier lieu une lemmatisation et un filtrage des requêtes, c'est à dire que les mots outils sont retirés et les mots fléchis sont ramenés à leur forme de base. Par exemple la phrase *Comment s'appellent le maires des Bastia* sera traitée à l'identique de *Comment s'appelle le maire de Bastia* via la phrase lemmatisée *Comment se appeller maire Bastia*. Ainsi nous proposons pour l'évaluation de comparer les versions lemmatisées des phrases de références et des phrases réécrites. De plus en accord avec un modèle étendu pour les hypothèses de la requête en entrée des systèmes, l'évaluation peut prendre en compte par exemple les trois premières hypothèses du système de réécriture, ou uniquement la première hypothèse. Les hypothèses fournies par le système de combinaison étant indépendantes, dans le cas de l'évaluation des trois premières hypothèses c'est le score de la phrase proposée la plus proche de la référence qui sera retournée.

Afin de comparer les performances de la combinaison phonétique et graphémique avec l'utilisation d'un correcteur orthographique pour la réécriture, nous avons évalué un système de réécriture basé uniquement sur les hypothèses fournies par Aspell, qui correspondent en réalité aux graphes de mots tel que celui de la figure 2. Dans ce cas, l'évaluation des trois premières hypothèses correspond à l'évaluation du chemin le plus proche de la référence.

Le tableau 3 contient les résultats de l'évaluation par SCLITE des phrases d'origine (Initial), des premières hypothèses du système graphémiques (Asp 1) et du système de combinaison (FSM 1), ainsi que des trois premières hypothèses fournies par ces systèmes (Asp 3 et FSM 3). L'évaluation est effectuée selon deux critères différents : le taux d'erreurs mots prend en compte à la fois les insertions, substitutions et omissions de mots, il faut le minimiser ; le pourcentage de phrases correctes constitue un aperçu des cas où l'on est certain que le système aura la possibilité de répondre, étant donné qu'il contient les même informations que la phrase de référence (il contient également des informations bruitées dans le cas où l'on considère plusieurs hypothèses de réécriture), il faut le maximiser. Les résultats pour les trois premières hypothèses

Mesure	Initial	Asp 1	Asp 3	FSM 1	FSM 3
Taux d'erreur	**51**	36	31	23	**20**
% phrases correctes	5	13	19	**43**	46

TAB. 3 – Taux d'erreur et pourcentage de phrases identiques à la référence après lemmatisation et filtrage sur les phrases tapées initialement ou réécrites à l'aide de Aspell (Asp) ou de notre système (FSM), si l'on considère la première ou les trois premières hypothèses.

montrent que si l'amélioration en terme de taux d'erreurs est déjà importante dans l'absolu (on le divise par 2,5 par rapport à l'initial), elle l'est aussi par rapport à un correcteur orthographique performant (le taux d'erreur est 1,5 fois plus bas). Les résultats en termes de taux d'erreurs sont

[4] `http://www.nist.gov/speech/tools`

également probants si l'on ne considère que la première hypothèse du système par combinaison, ce qui laisse à penser que l'ajout de bruit qu'apporterait des hypothèses multiples sera peut être plus néfaste que la perte engendrée par la conservation d'une seule hypothèse. Cela est confirmé par les résultats au niveau des phrases. En effet, on atteint 43 % de phrases identiques à la référence après filtrage et lemmatisation de la première hypothèse FSM, alors qu'il n'y en avait que 5 % à l'origine et qu'on atteint moins de 20 % avec Aspell. La différence avec l'évaluation des trois premières hypothèses FSM montre que la suggestion correcte se trouve en seconde ou troisième position seulement dans de rares cas.

Les résultats obtenus par les premières hypothèses du système de réécriture par combinaison sont très bons d'autant qu'il n'y a pas de dégradation des parties de phrases déjà correctes, et il est intéressant d'observer leur répartition en fonction des individus et des thèmes abordés (les thèmes étant les questions d'origine). Cette répartition consignée dans le tableau 4 montre que si les variations existent entre les individus, à part pour 1 et 4, elles ne sont pas très significatives étant donné qu'elles s'appliquent sur cinq exemples au maximum. La répartition des résultats par thème montre en revanche une nette différence, et l'on remarque notamment que les thèmes de questions 1 et 2 maintiennent des taux d'erreurs importants et que le système ne parvient à une phrase lemmatisée identique à la référence dans aucun cas. La raison de ces erreurs est que les noms propres associés à ces questions ne se trouvent ni dans le lexique phonétique ni dans le modèle de langage et sont par conséquent impossibles à proposer dans les hypothèses. Cela suggère que les performances du système par combinaison de processus graphémiques et phonétiques pourront encore être améliorés par un enrichissement dynamique des ressources, ou par un enrichissement statique se basant sur l'ensemble du corpus sur lesquelles les questions sont posées. En effet le modèle de langage a ici été appris sur un sous ensemble du corpus EQUER, et on peut imaginer y ajouter les phrases contenant des mots inconnus du lexique et du modèle initial.

Pers	Taux d'erreur initial	Taux d'erreur	% phrases correctes
1	36	9	75
2	67	39	40
3	49	18	40
4	50	30	20
5	73	12	60
6	30	27	40
7	46	21	40
8	60	30	50

Thème	Taux d'erreur initial	Taux d'erreur	% phrases correctes
1	58	47	0
2	52	42	0
3	40	0	100
4	43	17	62
5	65	11	57

TAB. 4 – Distribution des performances de notre système sur la première hypothèses proposée, par personne (Pers) ou par thème de question, selon les mesures de taux d'erreur mot et de pourcentage de phrases identiques à la correction, par rapport au taux d'erreur mot initial.

5 Conclusion

Les performances obtenues par un système combinant des aspects graphémiques et phonétiques au niveau de la phrase entière permettent de proposer des réécritures qui multiplient par 8 le nombre de questions correctes une fois lemmatisées, et donc d'autant les performances d'un système de questions réponses pour des questions tapées par des enfants dyslexiques. L'évaluation des phrases filtrées et lemmatisées montre que l'on peut faire descendre le taux d'erreurs

de 51 % à 23 % en considérant uniquement la première hypothèse, alors qu'un correcteur orthographique performant ne permet de descendre qu'à 36 %. Une analyse en profondeur des erreurs résiduelles montre qu'il est encore possible d'améliorer nettement les performances à l'aide de modèles de langages et de lexiques plus adaptés, soit plus complets soit dynamiques.

Références

ALLAUZEN C. & MOHRI M. (2005). The design principles and algorithms of a weighted grammar library. *International Journal of Foundations of Computer Science*, **16**(3), 403–421.

AYACHE C., GRAU B. & VILNAT A. (2006). Equer : the french evaluation campaign of question answering system equer/evalda. In *5th international Conference on Language Resources and Evaluation (LREC 2006)*, p. 1157–1160, Genoa, Italy.

BECHET F. (2001). Lia_phon - un systeme complet de phonetisation de textes. *Traitement Automatique des Langues (T.A.L.)*, **42**(1).

BRILL E. & MOORE R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the ACL*, p. 286–293.

DEOROWICZ S. & CIURA M. G. (2005). Correcting spelling errors by modelling their causes. *International journal of applied mathematics and computer science*, **15**(2), 275–285.

GILLARD L., BELLOT P. & EL-BÈZE M. (2005). Le lia à equer (campagne technolangue des systèmes questions-réponses). In *Actes de TALN'05*, Dourdan.

GILLARD L., SITBON L., BELLOT P. & EL-BEZE M. (2006). Dernieres evolutions de qualia, le systeme de questions/reponses du lia. *Traitement Automatique des Langues (TAL)*.

GILLON G. T. (2004). *Phonological Awareness- From Research to Practice*. Guilford Press.

JAMES A. & DRAFFAN E. (2004). The accuracy of electronic spell checkers for dyslexic learners. *PATOSS bulletin*.

LÉTÉ B., SPRENGER-CHAROLLES L. & COLÉ P. (2004). Manulex : A grade-level lexical database from french elementary-school readers. *Behavior Research Methods, Instruments, and Computers*, **36**, 156–166.

LOOSEMORE R. P. W. (1991). A neural net model of normal and dyslexic spelling. In *International Joint Conference on Neural Networks*, volume 2, p. 231–236, Seattle, USA.

MOHRI M., PEREIRA F. C. N. & RILEY M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, **16**(1), 69–88.

MOHRI M., PEREIRA F. C. N. & RILEY M. D. (1997). At&t fsm librarytm – finite-state machine library.

PEDLER J. (2001). The detection and correction of real-word spelling errors in dyslexic text. In *Proceedings of the 4th Annual CLUK Colloquium*.

SPOONER R. (1998). *A spelling checker for dyslexic users : user modelling for error recovery*. PhD thesis, Human Computer Interaction Group, Department of Computer Science, University of York, Heslington, York,.

TOUTANOVA K. & MOORE R. C. (2002). Pronunciation modeling for improved spelling correction. In *Proceedings of the 40th annual meeting of ACL*, p. 144–151, Philadelphia.

Vers une méthodologie générique de contrôle basée sur la combinaison de sources de jugement

Grégory SMITS[1], Christine CHARDENON[2]
[1] GREYC–Université de Caen, F-14032 CAEN cedex
[2] France Télécom R&D TECH/EASY/LN,
2, avenue Pierre Marzin, 22307 Lannion Cedex
gsmits@info.unicaen.fr
christine.chardenon@orange-ftgroup.com

Résumé. Le contrôle des hypothèses concurrentes générées par les différents modules qui peuvent intervenir dans des processus de TALN reste un enjeu important malgré de nombreuses avancées en terme de robustesse. Nous présentons dans cet article une méthodologie générique de contrôle exploitant des techniques issues de l'aide multicritère à la décision. À partir de l'ensemble des critères de comparaison disponibles et la formalisation des préférences d'un expert, l'approche proposée évalue la pertinence relative des différents objets linguistiques générés et conduit à la mise en place d'une action de contrôle appropriée telle que le filtrage, le classement, le tri ou la propagation.

Abstract. The control of concurrent hypotheses generated by the different modules which compose NLP processes is still an important issue despite advances concerning robustness. In this article, we present a generic methodology of control inspired from multicriteria decision aid methods. Based on available comparison criteria and formalized expert knowledge, the proposed approach evaluate the relevancy of each generated linguistic object and lead to the decision of an appropriate control action such as filtering, ordering, sorting or propagating.

Mots-clés : méthodologie de contrôle, aide multicritère à la décision, apprentissage automatique de métriques.

Keywords: control methodology, multicriteria decision aid, metrics automatic learning.

1 Introduction

De nombreuses avancées en termes de formalisme, d'algorithmique et de développement de ressources ont permis aux systèmes de traitement automatique des langues naturelles (TALN) d'atteindre une couverture très satisfaisante des différents phénomènes linguistiques observables. Cependant, quelle que soit la tâche ou le niveau d'analyse concerné, les différentes approches envisagées sont de manière récurrente confrontées au manque de précision des résultats générés. Ce phénomène se matérialise par la présence d'objets linguistiques concurrents de différentes natures. Bien que justifiable en présence d'ambiguïtés locales "naturelles", ces indéterminations concernent la plupart du temps des erreurs d'interprétations souvent qualifiées d'ambiguïtés "ar-

tificielles". Le contrôle du processus d'analyse a pour objectif d'identifier le plus tôt possible ces ambiguïtés "artificielles" afin notamment d'éviter leur propagation vers les étapes suivantes de l'analyse.

Bien que des propositions d'architectures de TALN conformes vis à vis des modèles cognitifs aient été proposées afin d'éviter la génération d'objets linguistiques erronnés (Rady, 1983) (Sabah, 1990), la mise en place de stratégies spécifiques de contrôle semble indispensable.

Nous verrons dans un premier temps que le contrôle des indéterminations, appelés "points d'embarras"[1] par (Sabah, 1989) repose sur la prise en compte d'informations distinctives hétérogènes. En s'appuyant sur un exemple de chaîne de TALN et un cas concret d'indétermination, nous verrons dans un second temps que le contrôle peut être considéré comme une démarche décisionnelle basée sur plusieurs critères de comparaison. Nous avons ainsi développé une méthode complète et générique inspirée des approches d'aide multicritère à la décision. Avant de présenter les premiers résultats obtenus, nous verrons que la prise d'une décision repose sur la formalisation des connaissances d'un expert sur la tâche à contrôler.

2 Définition de la notion de contrôle

2.1 Le TALN comme une succession de "points d'embarras" potentiels

De nombreux systèmes de TALN peuvent être représentés comme un ensemble de modules d'analyse qui appliqués successivement composent le processus complet d'interprétation linguistique. L'objectif de chacun de ces composants est de construire ses propres interprétations à partir de ressources linguistiques et des informations générées par les étapes précédentes. Ces interprétations que nous nommerons par la suite objets linguistiques peuvent correspondre à :
– des unités lexicales ou de sens ;
– à un découpage en constituants suite à une analyse syntaxique de surface ;
– à un arbre de dépendance syntaxique suite à l'application d'une grammaire de dépendances ;
– à des graphes sémantiques ;
– à des suggestions de corrections orthographiques ;
– à des rubriques sémantiques d'indexation correspondant à une requête.
Fréquemment appliqués séquentiellement, ces modules sont sources de "points d'embarras" et affectent donc la complexité et l'efficacité du processus d'analyse et nuisent également à l'utilisabilité des résultats générés. Il devient alors indispensable de contrôler l'apparition de ces indéterminations, notamment en comparant la pertinence relative de chaque objet linguistique instancié et en mettant en place une action de contrôle appropriée pour éviter la propagation d'interprétations erronnées.

2.2 Des "points d'embarras" aux "points de décision"

En conservant la terminologie proposée par (Sabah, 1989), l'enjeu du contrôle de ces processus réside dans la transformation des "points d'embarras" en "points de décision". Ce dernier désigne un état du processus de traitement où plusieurs objets linguistiques concurrents ont été générés, mais qui est également caractérisé par la disponibilité de connaissances et de critères de

[1] L'IA et le langage tome 2 page 121

comparaisons à partir desquels une stratégie de contrôle peut être déployée. Ainsi, pour pallier le manque d'informations distinctives nécessaires à la comparaison des différents objets linguistiques, de nombreux travaux ont proposé d'intégrer des connaissances supplémentaires de différentes natures (probabiliste (Blache & Rauzy, 2006), statistiques (Charniak & M.Johnson, 2005), heuristique (Uszkoreit, 1991), symbolique (Bourigault & Frérot, 2004)) pour qualifier la pertinence de chaque candidat et permettre la mise en place d'une action de contrôle.

On constate également que pour un "point d'embarras" identifié, plusieurs sources indépendantes de connaissances peuvent être exploitées pour affecter à chacun des objets linguistiques un critère de comparaison. Il apparaît cependant que pour chaque contexte de contrôle, aucune connaissance ne permet individuellement de caractériser et d'évaluer pleinement la validité de chaque interprétation envisagée.

L'indéterminisme lié à l'attachement d'un groupe prépositionnel constitue un exemple très illustratif . En effet, l'application d'une grammaire donnée sur une phrase "jouet" comme : "Je possède la statue de bois de rose de Charles", entraînerait sans doute la construction d'au moins 5 arbres syntaxiques concurrents. Afin de déterminer leur pertinence, chaque attachement peut être jugé par rapport :
– à sa conformité vis à vis des informations prosodiques (rarement disponibles) ;
– à son respect des heuristiques d'attachement droit ou minimal ;
– à des données de sous-catégorisations syntaxiques (Bourigault & Frérot, 2004) ;
– à sa fréquence d'apparition observée sur corpus (Gala, 2003) ;
– à des préférences d'usages liées à la sémantique des propositions (Whittemore *et al.*, 1990) ;
– etc.
Chacune des sources de jugement citées précédemment apporte ainsi une information discriminante permettant d'identifier pour certains cas "ambiguïtés" d'attachement envisageables les erreurs d'interprétation à filtrer, ou réciproquement les attachements à privilégier, mais également une erreur d'appréciation pour les autres cas. Afin d'augmenter la robustesse et la crédibilité apportée à l'évaluation de la pertinence de chaque objet linguistique, il semble indispensable de combiner et d'exploiter différents points de vue de jugement. Bien que des arguments psycholinguistiques (Altmann, 1998) (Gibson & Pearlmutter, 1998) aient déjà été avancés en faveur de l'usage combinée de sources de connaissances, peu de stratégies de contrôle exploitent la complémentarité de différents critères de comparaison (Rosso *et al.*, 2003) (Rigau *et al.*, 1997). Nos travaux se sont donc focalisés sur la mise en place d'un formalisme et d'une méthodologie générique de contrôle exploitant l'information apportée par chaque source de jugement pour comparer la pertinence d'objets linguistiques concurrents.

2.3 Le cas de `TiLT`

`TiLT` est une boîte à outils de TALN développée par l'équipe `Langues Naturelles` de `France Télécom R&D`. Le processus d'analyse paramétrable est composé d'un ensemble de modules de traitements qui, appliqués séquentiellement, construisent de manière itérative des interprétations linguistiques de différentes natures. En fonction des caractéristiques du contexte applicatif, certaines étapes de traitement constituent des "points d'embarras". Pour éviter la propagation d'objets linguistiques erronés, différents traits, scores ou probabilités calculés, méthodes spécifiques de jugement ont été intégrés au processus classique d'analyse pour être exploités comme critères de comparaison. Cependant aucune stratégie spécifique de contrôle ne permettait de les combiner ou d'évaluer leur efficacité. L'architecture modulaire d'analyse de

TiLT a été dans un premier temps modifiée (Smits, 2006) afin de simplifier l'intégration de ces critères de comparaison, de les centraliser et de permettre la mise en place de phases de contrôle entre les étapes d'analyse.

3 Formalisation d'une méthode d'aide multicritère à la décision

3.1 Agrégation des critères et interprétation des comparaisons

L'architecture décisionnelle de TiLT centralise en tant que critères de comparaison les connaissances supplémentaires intégrées lors du contrôle d'un "point d'embarras". En plus du cadre formel, l'approche de contrôle propose une méthode complète permettant à un expert, le linguiste ou l'informaticien en charge du paramétrage de TiLT, d'exprimer ses connaissances et intuitions sur la tâche de contrôle en question et la façon dont les critères disponibles doivent être exploités. Nous nous rapprochons ainsi du domaine de l'aide multicritère à la décision qui "vise, comme son nom l'indique, à fournir à un décideur des outils lui permettant de progresser dans la résolution d'un problème de décision où plusieurs points de vue, souvent contradictoires, doivent être pris en compte." (Vincke, 1998).

La méthode envisagée (Smits, 2007) s'inspire profondément des méthodes de surclassement (en particulier ELECTRE III et ELECTRE TRI (Roy, 1990)). À partir des valeurs des critères de comparaison qui qualifient les objets comparés et des connaissances exprimées par le décideur, ces méthodes établissent entre les différentes hypothèses candidates des relations de surclassement. Une telle relation notée $oSo\prime$ est établie entre deux objets linguistiques o et $o\prime$ si "il y a suffisamment d'arguments pour admettre que o est au moins aussi bonne que $o\prime$, sans qu'il y ait de raison importante de refuser cette affirmation." Ces relations "abstraites" de surclassements permettent d'établir des relations de préférence, d'indifférence (utile pour la factorisation) ou d'incomparabilité (exploitées pour le filtrage).

Pour répondre à une problématique de classement, on exploiter l'ensemble des relations de surclassement établies entre les différents objets linguistiques pour établir un pré-ordre partiel (avec *ex aequo*) ou total (Fig. 1). Quant aux problématiques de tri et filtrage, les différents objets linguistiques concurrents ne sont plus comparés entre eux, mais par rapport à des profils d'acceptabilité qui définissent, pour chaque classe considérée, les performances à atteindre sur chacun des critères pour faire partie de la classe en question (cf. cas d'expérimentation Fig. 2).

3.2 Formalisation du problème et construction des relations

Soit $O : o_1, o_2, ..., o_n$ l'ensemble des objets linguistiques comparés et $G : g_1, g_2, ..., g_m$ les m critères utilisés lors du contrôle, où $g_j(o_i)$ correspond à la valeur obtenue par l'objet o_i pour le j^{eme} critère.

Les préférences et connaissances du décideur se matérialisent dans un premier temps à travers les critères choisis pour la tâche de contrôle, mais également par un ensemble de paramètres qui peuvent être associés à chaque critère g_j :
– un poids d'importance w_j ;

– un seuil de préférence p_j ;
il correspond à la plus petite différence de valeur à partir de laquelle une situation de préférence peut être établie entre deux objets.
– un seuil d'indifférence q_j ($q_j <= p_j$) ;
correspond à la plus grande différence préservant l'indifférence entre 2 objets sur le critère j.
– un seuil de veto v_j ($p_j <= v_j$).
correspond à la différence de valeur à partir de laquelle un objet devient incomparable vis à vis d'un autre, car jugé trop faible sur un critère important. Ce paramètre permet notamment de définir des conditions de filtrage.

Ces préférences expertes interviennent dans le calcul d'un indice de surclassement $S(o, o\prime)$, quantifiant la crédibilité du surclassement de $o\prime$ par l'objet o, où $o\prime$ peut correspondre à un objet concurrent de o ou à un profil d'acceptabilité d'une classe pour les problématiques de tri. Cet indice de surclassement $S(o, o\prime)$ repose sur le produit d'un indice de concordance $C(o, o\prime)$, représentant la majorité des critères en faveur de o, et d'un indice de discordance $D(o, o\prime)$, représentant la minorité des critères refusant le surclassement de $o\prime$ par o : $S(o, o\prime) = C(o, o\prime)\ .\ D(o, o\prime)$, où $C(o, o\prime) = \frac{1}{\sum_{j \in G} w_j} \sum_{j \in G} w_j c_j(o, o\prime)$

$$c_j(o, o\prime) = \begin{cases} 1, & \text{si } g_j(o) - g_j(o\prime) >= p_j \\ 0, & \text{si } g_j(o) - g_j(o\prime) <= q_j \\ \frac{p_j - g_j(o) - g_j(o\prime)}{p_j - q_j}, & \text{si } q_j < g_j(o) - g_j(o\prime) < p_j \end{cases}$$

, et $D(o, o\prime) = \prod_{j \in \overline{G}} \frac{1 - d_j(o, o\prime)}{1 - C(o, o\prime)}$, $\overline{G} = j \in G / d_j(o, o\prime) > C(o, o\prime)$

$$d_j(o, o\prime) = \begin{cases} 1, & \text{si } g_j(o\prime) - g_j(o) >= v_j \\ 0, & \text{si } g_j(o\prime) - g_j(o) <= p_j \\ \frac{g_j(o\prime) - g_j(o) - p_j}{v_j - p_j}, & \text{si } p_j < g_j(o\prime) - g_j(o) < v_j \end{cases}$$

L'interprétation des relations de surclassement permet de définir des situations de préférence stricte : $oPo\prime$ si $S(o, o\prime) \wedge \neg S(o\prime, o)$, d'indifférence : $oIo\prime$ si $S(o, o\prime) \wedge S(o\prime, o)$ ou d'incomparabilité : $oRo\prime$ si $\neg S(o, o\prime) \wedge \neg S(o\prime, o)$. Les relations établies, regroupées dans une structure de préférences (Fig. 1), sont exploitées pour établir un pré-ordre complet ou partiel (avec *ex aequo*) des objets comparés.

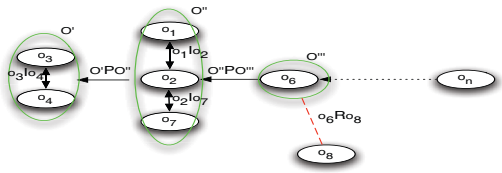


FIG. 1 – Représentation agrégée des relations de préférences établies entre les hypothèses

3.3 Vers une élicitation des préférences et connaissances de l'expert

La pertinence des relations de comparaison établies entre les objets concurrents dépend à la fois de l'information distinctive portée par les critères qui les qualifient, mais également des préférences du décideur. Il est cependant difficile et peu naturel pour un décideur de déterminer clairement et par l'intermédiaire de valeurs numériques, les valeurs de ces différents paramètres décisionnels. Même si l'externalisation de l'ensemble de ces informations dans un module spécifique de contrôle facilite les expérimentations itératives, l'impact de chaque paramètre sur le résultat final est difficilement quantifiable *a priori*. Il est en revanche nettement plus évident pour un expert de se prononcer sur la pertinence des objets linguistiques générés. Cette démarche de validation des résultats générés par un expert est fortement exploitée pour la construction de corpus de références, servant ensuite de données pour l'apprentissage de ressources ou l'évaluation de systèmes.

Nous proposons de considérer l'identification des objets linguistiques de référence comme l'expression de l'expertise de l'annoteur sur une tâche de contrôle à automatiser. Ainsi, qualifier une hypothèse comme valide revient à qualifier les performances qu'elle a obtenu sur les critères exploitées comme discriminants. Une hypothèse annotée comme correcte ou incorrect et les performances calculées sur les critères concernés par l'étape de contrôle constituent un enregistrement de ce que nous nommons un tableau de performances : Nous exploitons ce tableau de

Objets annotés.	vecteur de performances				annotation
	critère 1	critère 2	...	critère m	
o_1	4.2	Vrai	...	36	correct
o_2	5.0	Vrai	...	16	correct
o_3	2.6	Faux	...	24	incorrect
o_4	1.2	Faux	...	42	correct
...	...	...	...	...	...
o_{p-1}	4.0	Vrai	...	4	incorrect
o_p	0	Faux	...	17	incorrect

TAB. 1 – Tableau de performances construit à partir du corpus de références

performances pour évaluer et quantifier la pertinence de chacun des critères disponibles. Une distribution de performances d'un critère est jugée pertinente, si elle permet de caractériser un certain type d'annotation (i.e. la classe des hypothèses correctes ou la classes des hypothèses incorrectes). La méthode d'apprentissage de métriques `RELIEF` (Kononenko, 1994) permet d'atteindre ce but, en associant à chaque critère un poids normé sur $[-1, 1]$, où une valeur négative caractérise un critère non représentatif de la classe des hypothèses correctes. Les résultats de la méthode sont ensuite exploiter lors de la construction de l'indice de crédibilité en tant que vecteur de poids des critères.

Nous proposons également de considérer la performance minimale obtenue sur un critère par les hypothèses annotées comme correctes en tant que limite d'acceptabilité de ce critère.

4 Expérimentation

Nous présentons dans cette section les premiers résultats obtenus sur un des nombreux cas de contrôle envisageables. Il s'agit de répondre à une problématique de classement des couples antécédent/reprise-anaphorique candidats extraits d'un corpus.

4.1 Contrôle des couples antécédent/reprise-anaphorique candidats

Cette expérimentation s'inscrit dans le cadre d'une collaboration et d'une extension des travaux réalisés par OLIVIER TARDIF (Tardif, 2006). Un algorithme extrait à partir d'un texte un ensemble de couples d'expressions constituant potentiellement des patrons antécédent/reprise-anaphorique. Une expression correspond à une entité nommée (NPR : Mickaël Gordbatchev, l'URSS, Vilnius), un pronom (PRON : il, celui-là) ou un groupe nominal (NCOM : le dirigeant soviétique, le parlement). L'enjeu du contrôle est de construire une classes des candidats valides à partir des performances qui leurs sont associées sur différents critères, tels que :
– des mesures de distances (en mots, phrases, etc.) ;
– la correspondance des classes sémantiques et des fonctions syntaxiques ;
– des marques morphologiques (indéfini, possessif) ;
– des mesures de distance et de similarité alphabétiques ;
– des propriétés d'accords de genre et de nombre.
Pour constituer la classe des reprises anaphoriques valides, les relations de surclassement ne sont pas construites entre les couples candidats, mais entre chaque candidat et un profil d'acceptabilité (Fig. 2). Ce vecteur de limites d'acceptabilité constitue une nouvelle préférence mise en place par l'expert pour contrôler la validité des hypothèses comparées. Ainsi, un candidat qui surclasse ce profil est considéré comme un cas de reprise anaphorique.

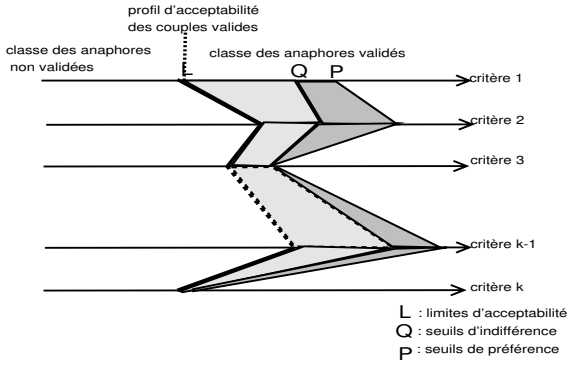


FIG. 2 – Profil d'acceptabilité des reprises anaphoriques candidates

4.2 Entre expertise et apprentissage automatique

Nous disposons d'un corpus de 80 textes journalistiques (`Le Monde` de 1989-1990) annoté automatiquement par `TiLT` afin de disposer d'informations morphologiques, syntaxiques et sémantiques sur les expressions et leur rôle dans la phrase. Les liens de coréférences entre les expressions ainsi que la classe sémantique (personne, lieux, organisation) de celles-ci ont ensuite été marqués manuellement.

Les couples d'expressions constituées à partir du corpus ont été partitionnés en fonction de leur type : NPR-NPR, NPR-NCOM, NPR-PRON, NCOM-NCOM, NCOM-PRON, PRON-PRON. Nous nous sommes restreint pour cette évaluation aux couples ayant pour antécédent un nom propre.

Dans un premier temps, nous avons, à travers une démarche interactive, demandé à un expert du domaine de constituer trois profils de paramètres décisionnels (voir Sec. 3.2) pour les trois cas de reprise anaphorique traités. L'expert devait ainsi identifier les critères qu'il jugeait pertinents dans chacun des cas, ainsi que leur importance relative dans l'agrégation, un profil d'acceptabilité et éventuellement des seuils de préférence, indifférence et veto. Dans un second temps, nous avons exploité une partie du corpus annoté pour apprendre automatiquement les poids des différents critères ainsi que les seuils délimitant la classe des couples valides (Sec. 3.3). Ce corpus est constitué de 950 paires npr-npr, 3400 paires npr-ncom et 620 paires npr-pron, composé respectivement de 90, 46 et 48 cas valides (positifs) de coréférence. Nous avons ensuite évalué les différents profils de paramètres sur un corpus d'évaluation extrait du corpus de référence, constitué de 120 paires candidates NPR-NPR (19 positives), 80 paires NPR-PRON (12 positives) et de 414 paires NPR-NCOM (9 positives).

Profil	Expressions	Précision	Rappel	F-mesure
Manuel	NPR-NPR	0.9	0.92	0.91
	NPR-NCOM	0.4	0.2	0.26
	NPR-PRON	0.4	0.25	0.3
Automatique	NPR-NPR	0.94	0.96	0.95
	NPR-NCOM	0.35	0.17	0.23
	NPR-PRON	0.6	0.39	0.47

L'apprentissage automatique des poids des critères ainsi que des limites d'acceptabilité nous permet d'améliorer sensiblement les résultats bien que le corpus soit principalement composé d'exemples négatifs (à plus de 90% sur le corpus de test et à plus de 95% sur le corpus d'apprentissage). Inférer automatiquement ces paramètres décisionnels nous permet identifier et de quantifier l'utilité des différents critères disponibles, contredisant parfois les intuitions de l'expert qui exploitait des critères n'apportant que du bruit. Par exemple, pour le cas des couples NPR-PRON, l'expert a sélectionné quatre critères comme pertinents et a formulé les préférences suivantes concernant l'importance relative de chacun d'eux :

1. accord en nombre entre l'antécédent et la reprise
2. accord en genre entre l'antécédent et la reprise
3. nombre d'occurences de l'antécédent dans le texte
4. nombre d'expressions séparant l'antécédent de la reprise
5. antécédent et reprise ont la fonction sujet

Cependant, par apprentissage sur corpus de référence, de nouveaux critères ont été identifiés comme pertinents et l'ordre d'importance de l'ensemble des critères utilisé a été modifié, ce qui explique l'amélioration des résultats (les autres paramètres de seuils restants identiques au profil décisionnel de l'expert) :

1. antécédent est l'expression la plus proche
2. antécédent et reprise sont dans la même phrase
3. nombre d'occurences de l'antécédent dans le texte
4. accord en nombre entre l'antécédent et la reprise
5. nombre d'expressions séparant l'antécédent de la reprise
6. nombre de mots séparant l'antécédent de la reprise
7. antécédent et reprise ont la fonction sujet
8. accord en genre entre l'antécédent et la reprise

Cette tâche d'identification des couples antécédent/reprise-anaphorique avait dans un premier été traitée à l'aide de classifieurs bayésiens naïfs. Outre des améliorations des valeurs de précision et de rappel, notre approche offre à l'expert la possibilité d'intervenir sur le comportement de la méthode de classification mais également une meilleure compréhension des décisions émises.

5 Perspectives et conclusion

Nous proposons une méthode générique de contrôle des points d'embarras apparaissant lors d'un processus de TALN. Cette méthode inspirée de l'aide multicritère à la décision se base sur l'agrégation de critères de comparaison hétérogènes. Les différents paramètres externalisés dans un profil décisionnel permettent à un expert d'exprimer ses connaissances et intuitions sur le problème traité. Pour valider ou inférer automatiquement les préférences émises par un expert, nous utilisons des méthodes d'apprentissage supervisé exploitant un corpus de référence.

La méthode d'apprentissage de métriques `RELIEF` s'avère efficace pour quantifier la représentativité d'un critère vis à vis d'un ensemble d'exemples annotés comme valides. Nous envisageons cependant d'exploiter une variante de cette méthode pour réduire l'impact de la forte proportion d'exemples négatifs lors de l'apprentissage. Nous travaillons actuellement sur la mise en place de méthodes de seuillage pour inférer automatiquement les autres paramètres décisionnels.

L'approche proposée est en cours de validation sur un autre cas concret d'expérimentation : le classement des arbres syntaxiques concurrents. L'automatisation d'une procédure d'aide à la décision basée sur la comparaison deux à deux des hypothèses concurrentes, pour les problématiques de classement, pose cependant des problèmes de complexité. Ainsi, en présence d'un grand nombre d'hypothèses concurrentes, il ne semble pas judicieux de construire un classement de tous les candidats. Nous proposons donc d'effectuer un premier filtrage en exploitant notamment les critères jugés comme les plus pertinent par la méthode d'apprentissage automatique des poids. Sur le sous-ensemble d'hypothèses restant, des relations de surclassement peuvent être établies et interprétées pour obtenir un classement des N meilleurs candidats.

Références

ALTMANN G. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, **2**(4).

BLACHE P. & RAUZY S. (2006). Mécanismes de contrôle pour l'analyse en grammaires de propriétés. In *in proceedings of TALN*.

BOURIGAULT D. & FRÉROT C. (2004). Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène. In *in proceedings of TALN*.

CHARNIAK E. & M.JOHNSON (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *ACL '05 : Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 173–180, Morristown, NJ, USA : Association for Computational Linguistics.

GALA N. (2003). Une méthode non supervisée d'apprentissage sur le web pour la résolution d'ambiguïtés structurelles liées au rattachement prépositionnel. In *in proceedings of TALN*.

GIBSON E. & PEARLMUTTER N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences*, **2**(7).

KONONENKO I. (1994). Estimating attributes : Analysis and extensions of relief. In *In proceedings of the European Conference on Machine Learning*.

RADY M. (1983). *L'ambiguïté du langage naturel est-elle la source du non-déterminisme des procédures de traitement ?* PhD thesis, Université de Paris VI.

RIGAU G., ATSERIAS J. & AGIRRE E. (1997). Combining unsupervised lexical knowledge methods for word sense disambiguation. *Proceedings of the 35th annual meeting on Association for Computational Linguistic*.

ROSSO P., MASSULLI F. & BUSCALDI D. (2003). Word sense disambiguation combining conceptual distance, frequency and gloss. *IEEE*.

ROY B. (1990). Decision-aid and decision-making. In *European Journal of Operational Research*, volume 45, p. 324–331.

SABAH G. (1989). *L'IA et le langage (tome 2)*. Hermes.

SABAH G. (1990). Caramel : A flexible model for interaction between the cognitive processes underlying natural language understanding. In *Proceedings of the Ninth European Conference on Artificial*.

SMITS G. (2006). Contrôle dynamique multicritère des résultats d'une chaîne de tal. In *in proceedings of RECITAL*.

SMITS G. (2007). Méthodologie d'aide multicritère à la décision pour le contrôle d'une chaîne de traitement automatique des langues naturelles. In *in proceedings of ROADEF'07*.

TARDIF O. (2006). Résoudre la coréférence à l'aide d'un classifieur bayésien naïf. In *in proceedings of RECITAL*.

USZKOREIT H. (1991). Strategies for adding control information to declarative grammars. In A. FOR COMPUTATIONAL LINGUISTICS, Ed., *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, p. 237–245.

VINCKE P. (1998). *Aide multicritère à la décision*. Ellipses Marketing.

WHITTEMORE G., FERRARA K. & BRUNNER H. (1990). Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*.

Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques

Agnès TUTIN
LIDILEM, Université Grenoble 3, BP 25, 38040 Grenoble Cedex 09
agnes.tutin@u-grenoble3.fr

Résumé. Dans cette étude sur le lexique transdisciplinaire des écrits scientifiques, nous souhaitons évaluer dans quelle mesure les méthodes distributionnelles de TAL peuvent faciliter la tâche du linguiste dans le traitement sémantique de ce lexique. Après avoir défini le champ lexical et les corpus exploités, nous testons plusieurs méthodes basées sur des dépendances syntaxiques et observons les proximités sémantiques et les classes établies. L'hypothèse que certaines relations syntaxiques - en particulier les relations de sous-catégorisation – sont plus appropriées pour établir des classements sémantiques n'apparaît qu'en partie vérifiée. Si les relations de sous-catégorisation génèrent des proximités sémantiques entre les mots de meilleure qualité, cela ne semble pas le cas pour la classification par voisinage.

Abstract. In this study about general scientific lexicon, we aim at evaluating to what extent distributional methods in NLP can enhance the linguist's task in the semantic treatment. After a definition of our lexical field and a presentation of our corpora, we evaluate several methods based on syntactic dependencies for establishing semantic similarities and semantic classes. Our hypothesis that some syntactic relations – namely subcategorized relations – is more relevant to establish semantic classes does not entirely appears valid. If subcategorized relations produce better semantic links between words, this is not the case with neighbour joigning clustering method.

Mots-clés : corpus – écrits scientifiques - classes sémantiques – analyse distributionnelle.

Keywords: corpus – scientific writings – semantic classes – distributional analysis.

1 Introduction

Le traitement sémantique des éléments du lexique constitue un préalable dans de nombreuses applications du TAL. Dans une application d'aide à la rédaction en Français Langue Etrangère (Kraif & Tutin 2006), nous souhaitons ainsi effectuer un traitement du lexique transdisciplinaire des écrits scientifiques et de ses collocations. Dans ce cadre, nous souhaiterions proposer une approche onomasiologique de ce lexique (i.e. avec un accès par le sens plutôt que par la forme), dont l'étude pourrait être facilitée si les approches « machinales » (Habert & Zweigenbaum 2003) de traitement sémantique à partir d'analyse distributionnelle se révélaient concluantes pour le travail du linguiste. Dans cette étude, nous désirons plus précisément évaluer la pertinence des méthodes d'analyse distributionnelle basées sur des dépendances syntaxiques pour la constitution de classes sémantiques homogènes de noms transdisciplinaires des écrits scientifiques. Nous voudrions en particulier déterminer dans quelle mesure cette méthode, qui s'est révélée adaptée à des sous-langages spécifiques pour la terminologie du droit (Bourigault & Lame 2002), de l'immunologie (Harris *et al*. 1989) ou de la médecine (Nazarenko *et al.* (2001), peut être appliquée au lexique du genre des écrits scientifiques qui présente davantage de polysémie. Nous faisons l'hypothèse

que certaines relations syntaxiques de dépendance, plus contraintes sur le plan syntaxique et sémantique, produiront des associations sémantiques de meilleure qualité.

Dans un premier temps, nous définirons le lexique transdisciplinaire des écrits scientifiques, et présenterons un premier classement sémantique manuel basé sur des propriétés linguistiques. Dans un second temps, nous évaluerons les résultats de la méthode distributionnelle employée par Didier Bourigault (Bourigault 2002 ; Bourigault et Lame 2002) à notre lexique, méthode qui dissocie les « voisins en tête » des « voisins en expansion », et les comparerons au classement manuel. Puis, nous nous pencherons sur une seconde méthode basée sur les dépendances syntaxiques que le mot soit recteur ou régi (à l'instar de Grefenstette 1996). Nous comparerons enfin les associations établies avec les relations syntaxiques de sous-catégorisation et les associations issues des relations de modification. Nous finirons par une évaluation et une réflexion sur les méthodes distributionnelles « machinales » pour la tâche linguistique qui nous intéresse.

2 Le lexique transdisciplinaire des écrits scientifiques : un premier classement manuel

Le lexique transdisciplinaire des écrits scientifiques, qui apparaît dans les articles de recherche, les monographies scientifiques, les mémoires, les thèses et les rapports de recherche, est le lexique partagé par la communauté scientifique mis en œuvre dans la description et la présentation de l'activité scientifique. Ce lexique peut être considéré comme un lexique de genre, n'intégrant pas la terminologie du domaine, mais renvoyant aux concepts mis en œuvre dans l'activité scientifique (*examiner, prouver, réfuter, concluant, hypothèse, examen, encourageant ...*) (Cf. aussi les définitions un peu différentes du VGOS de Phal (1971) et les travaux de Pecman (2004) sur le lexique des écrits des sciences « dures »). Nous nous intéressons en particulier au lexique méthodologique partagé par l'ensemble des disciplines scientifiques, qu'il s'agisse des sciences expérimentales, des sciences appliquées ou des sciences humaines.

L'étude de ce lexique permet d'approfondir au plan linguistique et épistémologique la spécificité de l'écrit scientifique en repérant un ensemble de traces lexicales emblématiques du genre. Ce traitement peut également déboucher sur des applications didactiques comme l'aide à la rédaction en langue maternelle et en langue étrangère. Dans cette perspective, nous souhaiterions proposer des outils facilitant le choix lexical pour les apprenants étrangers, basés sur un accès onomasiologique (accès par l'analogie ou la classe sémantique) ou sémasiologique (par la forme) (Cf. Kraif & Tutin 2006). A cet effet, un premier relevé basé sur les noms fréquents et communs à des corpus de plusieurs disciplines a été effectué puis filtré[1]. Dans un second temps, ces noms ont été répartis dans des grandes classes sémantiques, à partir de propriétés syntaxiques, morphologiques et sémantiques, un peu à la façon de Flaux et van de Velde (2000) pour les noms abstraits. Pour les 83 noms les plus fréquents, sept grandes classes ont été dégagées :

[1] Ont été retenus un ensemble de noms (catégorisation de Cordial) apparaissant plus de 15 fois en médecine, linguistique et économie dans un corpus de 2 millions de mots.

1 **Les noms de processus de l'activité scientifique** (*analyse, application, choix, ...*) sont des noms extensifs (se combinent avec *lors, durant*, des verbes phasiques, souvent avec *faire*), et ont un agent humain.
2 **Les noms d'objets construits par l'activité scientifique** (*approche, argument, concept, conception, démarche, ...*) ne sont pas extensifs, ont un agent humain, se combinent avec des verbes comme *élaborer, construire*.
3 **Les noms d'observables de l'activité scientifique** (*cas, données, échantillon, exemple, facteur,...*) ne sont pas extensifs, se combinent avec le support *être* et avec les verbes *analyser, examiner, étudier*.
4 **Les noms de supports de la rédaction scientifique** (*article, chapitre, conclusion, document, figure, ...*) sont à la fois concrets et abstraits non extensifs. Ils se combinent avec la préposition *dans*, et sont sujets du verbe *présenter*.
5 **Les noms de caractérisation** (*caractère, caractéristique, différence, difficulté, fonction,…*) sont des noms intensifs, se combinant souvent avec le support *avoir* et sont généralement accompagnés d'un adjectif.
6 **Les noms d'acteurs de l'activité scientifique** (*auteur, chercheur, …*) sont des noms humains, souvent sujets des verbes d'activité scientifique (*examiner, décrire, observer …*).
7 **Les noms de relation logique** (*but, cause, conséquence, corrélation, effet, influence, liaison, lien, rapport, relation…*), qui sont abstraits et non extensifs, se combinent avec les supports *être* et *avoir* et apparaissent souvent dans des structures : Nlogique de N..

Les noms polysémiques comme *rapport* ou *étude* sont bien entendu rattachés à plusieurs classes. Ce premier classement sera notre étalon pour l'évaluation des méthodes distributionnelles automatiques.

3 Le corpus des écrits scientifiques

Les méthodes distributionnelles machinales sont tributaires des données textuelles exploitées. La qualité des associations lexicales extraites dépend en effet très largement de l'homogénéité et de la représentativité des corpus traités. Pour cette étude, nous avons constitué un corpus de 2 millions de mots comprenant plusieurs genres d'écrits scientifiques du français (articles scientifiques, thèses, rapports, cours) dans trois disciplines assez différentes : la linguistique, l'économie et la médecine (Le tableau 1 indique le nombre de mots pour chaque type de texte). Le corpus d'articles scientifiques est extrait du corpus KIAP[2] élaboré par l'équipe de Kjersti Fløttum, de l'Université de Bergen. Notre objectif sera d'observer comment s'effectuent les regroupements des noms transdisciplinaires qui ont des comportements syntaxiques analogues.

	Linguistique	**Economie**	**Médecine**
Articles de revues (corpus KIAP)	285 881 mots	374 516 mots	164 315 mots
Thèses, rapports, cours	364 812 mots	286 653 mots	492 173 mots
Total	650 693 mots	661 169 mots	656 488 mots

Tableau 1 : Corpus des écrits scientifiques

[2] KIAP : Kulturell Identitet i Akademisk Prosa. Cf. http://kiap.aksis.uib.no/

4 La méthode distributionnelle du linguiste et l'analyse distributionnelle machinale

Pour établir des associations sémantiques, l'intérêt de l'analyse distributionnelle paraît aller de soi, puisqu'il est classique dans la tradition de la sémantique lexicale, en particulier européenne (Cruse 1986, par exemple), de considérer que des mots qui ont des environnements syntaxiques comparables partagent des propriétés sémantiques non triviales, allant de la synonymie pour les associations les plus fortes à la co-hyponymie (Cf. aussi l'étude réalisée par Galy & Bourigault (à paraître)). Le recours aux distributions syntaxiques pour mettre en évidence les propriétés sémantiques permet au linguiste de s'appuyer sur des critères tangibles, palpables, et non plus des approximations notionnelles.

Cependant, la méthode distributionnelle du linguiste, qui fait appel en partie à son intuition de sujet parlant et catégorisant, diffère assez largement de l'approche distributionnelle « orthodoxe », en particulier dans Harris et al. (1989), entièrement basée sur les observables du corpus. En effet, le linguiste choisit tout d'abord les contextes lexicaux qui lui apparaissent les plus pertinents pour circonscrire la notion qui l'intéresse (Cf. par exemple, la notion de classe d'objets, chez Gaston Gross (1994)). Dans notre champ lexical, par exemple, on pourra ainsi repérer comme 'objets construits par l'activité scientifique' des noms qui se combinent régulièrement avec les verbes de la série *élaborer, construire, concevoir*. Le linguiste laissera de côté les associations lexicales qui lui apparaissent moins déterminantes, contrairement à l'approche automatique qui ne peut pas sélectionner *a priori* les contextes lexicaux qui seront les plus révélateurs. En outre, le linguiste effectue naturellement la désambiguïsation des notions, par exemple *conclusion* comme partie du texte, ou comme aboutissement d'un raisonnement, opération qui sera beaucoup plus délicate avec une méthode machinale. Enfin, le linguiste complète les données lacunaires du corpus ou écarte les associations jugées atypiques. Si un contexte n'est pas observable dans les textes, il recourt à son intuition pour vérifier si le contexte est possible. En bref, le linguiste s'aide du corpus, mais s'en abstrait partiellement pour les besoins interprétatifs si besoin est.

La méthode distributionnelle « orthodoxe » apparaît plus contrainte, puisqu'elle doit permettre de tirer toutes les observations du corpus et rien que du corpus. Le corpus doit donc à la fois être exhaustif pour la représentativité des associations lexicales (donc de grande taille), et très homogène pour éviter la polysémie. Cette approche donne généralement de bons résultats dans le domaine de la terminologie (Bourigault & Lame 2002 ; Harris *et al*. 1989 ; Nazarenko *et al.* 2001) où le lexique présente peu de variations. Nous souhaitons évaluer la même méthode dans notre champ lexical, en exploitant des relations syntaxiques de dépendance. Nous faisons l'hypothèse qu'en sélectionnant certains types de relation, à l'instar de la méthode distributionnelle « manuelle », nous obtiendrons des résultats de meilleure qualité.

5 Évaluation de méthodes d'analyse distributionnelle machinale basées sur des dépendances syntaxiques

Dans les méthodes d'analyse distributionnelle machinale, plusieurs définitions de la distribution ont été proposées. Les plus rustiques (Cf. par exemple Grefenstette (1996) peuvent simplement prendre en compte les mots pleins partagés dans une fenêtre de quelques mots. Les distributions basées sur les relations syntaxiques partagées donnent cependant de meilleurs résultats sur les lexèmes les plus fréquents, donc les plus significatifs (Grefenstette Ibid.). Nous adopterons cette dernière méthode en exploitant les dépendances syntaxiques obtenues sur

notre corpus à l'aide des résultats de l'analyseur Syntex (Bourigault *et al.* 2005). Nous évaluerons les proximités sémantiques établies et les classes sémantiques obtenues à l'aide des coefficients de similarité entre les mots.

5.1 Proximités sémantiques établies à l'aide de la méthode de D. Bourigault (2002)

La méthode distributionnelle a été appliquée avec succès par Didier Bourigault et ses collègues à plusieurs domaines dont la terminologie du droit (Bourigault & Lame 2002). Cette approche présente deux originalités : d'une part, elle dissocie les mots proches, appelés « voisins », selon qu'ils sont recteurs (ou têtes) ou régis (dans l'expansion) ; d'autre part, comme elle vise les applications terminologiques, elle prend en compte aussi bien les unités que les syntagmes dans les relations syntaxiques de dépendance. Le système, appelé Upery, basé sur les résultats de l'analyse syntaxique du logiciel Syntex (Bourigault et *al.* 2005), extrait des triplets contenant le terme (unité lexicale simple ou complexe), la relation de dépendance, et le contexte (le syntagme ou élément lexical régi). Il rapproche ensuite, en utilisant des mesures de proximité comme le jaccard, les termes selon le nombre de contextes différents qu'ils partagent[3]. Par exemple, les mots *article* et *chapitre*, qui apparaissent à la première ligne du tableau, partagent 6 contextes identiques (= a) lorsqu'ils sont accompagnés d'un adjectif[4] (par exemple *présent, suivant, dernier ...*). *article* apparaît lui-même dans 18 contextes adjectivaux différents (= n1), alors que *chapitre* apparaît lui-même dans 12 contextes adjectivaux (= n2). Le coefficient jaccard utilisé ici calcule la proximité sémantique entre les mots avec la formule suivante : $a/(n_1+n_2-a)$. Seuls sont sélectionnés les voisins pour lesquels le coefficient de jaccard dépasse 0,10 et qui ont au moins quatre types de contextes communs.

contexte1	rel1	contexte2	rel2	a	n1	n2	jaccard
article	ADJ	chapitre	ADJ	6	18	12	0.25
article	EPI	section	EPI	6	11	19	0.25
tableau	EPI	chapitre	EPI	21	84	21	0.25

Tableau 2 : Exemples de voisins en tête extraits à l'aide de l'outil Upery de Didier Bourigault

Upery a été appliqué à notre corpus d'écrits scientifiques et sur le lexique des 85 noms transdisciplinaires classés. Nous avons ensuite évalué les couples extraits à partir des classes établies manuellement, en examinant tour à tour les voisins en expansion et les voisins en tête.

Les voisins en tête associent des mots qui sont des recteurs et qui partagent des contextes semblables avec une relation syntaxique donnée. Pour la liste de noms sélectionnés, on obtient 516 résultats. Nous avons observé pour chaque couple de voisins établi si les deux éléments associés appartenaient à la même classe dans notre classification manuelle. Si tel était le cas, nous avons considéré que la réponse était acceptable et l'avons rejetée dans le cas inverse. Par exemple, l'association *figure-chapitre* a été considérée comme satisfaisante car les deux noms font partie de la classe des 'supports écrits de l'activité scientifique', mais l'association

[3] La méthode ne prend pas en compte le nombre d'occurrences pour chaque contexte, contrairement à d'autres approches comme celle de Grefenstette (1996) mais seuls sont retenus les contextes apparaissant plus de deux fois.

[4] Les relations pourraient ici être différentes pour les deux éléments rapprochés.

hypothèse-section n'apparaît pas valide car les deux éléments appartiennent à des classes différentes.

L'observation des résultats révèle que 50,5 % des voisins en tête extraits relèvent de la même classe, ce qui est *a priori* assez peu, étant donné le caractère assez lâche des classes établies manuellement. Les voisins en tête mettent en jeu de nombreuses relations de modification[5], facultatives, et peu contraintes sur le plan sémantique, comme la relation d'épithète ou d'attribut. Par exemple, les noms *cas* et *modèle*, assez distincts sur le plan sémantique, apparaissent dans 19 contextes adjectivaux communs. Un examen plus poussé montre que nombre de ces adjectifs sont très peu contraints du point de vue de leur sélection nominale (par exemple, *autre, dernier, tel, général, précédent*) et donc probablement peu informatifs du point de vue sémantique.

Nous avons ensuite comparé ces résultats avec les voisins en expansion, c'est-à-dire les cas où les noms transdisciplinaires sont régis dans une relation de sujet ou de complément. Nous faisons l'hypothèse que ces relations qui mettent souvent en jeu des arguments sous-catégorisés – mais pas uniquement –, souvent obligatoires, seraient davantage significatives pour établir des proximités sémantiques. Les résultats obtenus, bien que peu nombreux, semblent aller dans ce sens. Utilisant les mêmes seuils que pour les voisins en tête, 52 paires de voisins sont dégagées, dont 34 apparaissent valides (65,5% des paires). L'examen plus détaillé des contextes partagés montre que les associations Nom-Verbe apparaissent souvent plus significatives que dans les contextes Nom-Adj, à l'exception des relations où le verbe *être* apparaît.

5.2 Proximités sémantiques établies à l'aide de l'ensemble des relations syntaxiques

La méthode de Didier Bourigault dissocie les voisins qui apparaissent comme têtes des voisins qui apparaissent comme régis (dans l'« expansion »). Ce traitement séparé permet de mettre en lumière des associations spécifiques, comme l'association *examiner des données* et *l'examen des données*, qui seraient autrement noyées dans l'ensemble des relations. Ce type d'observation n'étant pas essentiel pour notre étude, nous avons observé, à l'instar de Grefenstette (1996) les proximités sémantiques établies à partir de l'ensemble des relations syntaxiques, que le nom transdisciplinaire soit recteur (Ex : *analyse des données*) ou régi (Ex : *confirmer l'analyse* ...). Les contextes ont ici été réduits aux verbes, noms et adjectifs qui entretenaient une relation syntaxique avec le nom transdisciplinaire, et non plus à tous les éléments (mots simples ou syntagmes) apparaissant en cooccurrence. L'idée était ici de vérifier si une fusion des relations, en produisant un plus grand nombre de contextes communs, pouvait améliorer la qualité des résultats.

La méthode employée (avec les mêmes seuils qu'en 5.1) produit 292 paires, dont 177 (soit 60,5%) apparaissent correctement appariées. Les résultats apparaissent donc meilleurs que pour les voisins en tête, mais cependant inférieurs à ceux des voisins en expansion.

[5] Mais pas uniquement. On repère aussi des relations de compléments de noms comme dans *l'efficacité de cette méthode* ou *l'élaboration du modèle*.

En outre, une classification par voisinage (neighbour joining cluster) a été effectuée à partir d'une une matrice contenant tous les coefficients de proximité (jaccard) – sans seuil – liant les mots (Cf. Fig. 1.a). Sur les 27 classes finales obtenues, 20 constituent des sous-ensembles des 7 classes définies manuellement (2 sous-ensembles ont des éléments uniques). Les sous-classes révèlent des associations lexicales fines, qui apparaissent pour la plupart appropriées pour notre approche onomasiologique.

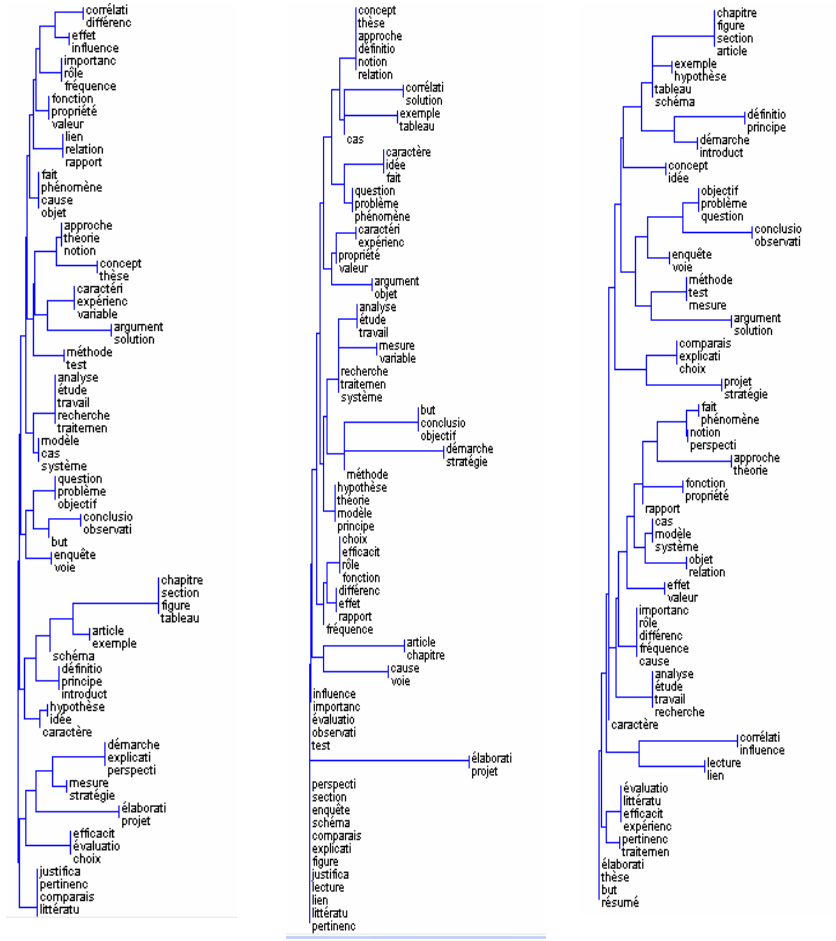
5.3 Proximités sémantiques établies à l'aide des relations de sous-catégorisation *vs* relations de modification

Nous faisons l'hypothèse que les relations syntaxiques mettant en jeu la sous-catégorisation sont plus déterminantes pour établir des proximités sémantiques que les relations de modification, parce que les arguments sont davantage contraints sur le plan syntaxique et sémantique par les restrictions sélectionnelles. Les voisins en expansion - correspondant pour la plupart à des relations de sous-catégorisation - obtenus avec la méthode de Didier Bourigault semblaient aller dans ce sens. Nous avons souhaité approfondir ce point en observant plus systématiquement quelques relations de sous-catégorisation. Les relations de sous-catégorisation observées ont été la relation objet (*confirmer une analyse*), la relation sujet (*les résultats infirment* ...), les compléments nominaux en *de*, que le nom soit recteur (ou tête) (*l'analyse des données*) ou régi (*l'efficacité de la méthode*)[6].

La méthode a dégagé 76 paires, dont 48 ont été considérées valides, soit 63 %. Nous avons ensuite comparé ces résultats avec les associations obtenues uniquement avec les modifieurs. Pour cela, nous avons sélectionné uniquement les relations liant l'adjectif épithète au nom, ainsi que la relation d'apposition. 582 paires ont été obtenues, parmi lesquelles 285 ont été validées, soit 49%. On remarque donc que le nombre de paires obtenues par les relations de sous-catégorisation apparaît nettement moins important que le nombre de paires obtenues à l'aide des relations de modification. Cette disparité des effectifs semble avoir une incidence sur les classes établies à l'aide de la même méthode qu'en 5.2 (Cf. Fig. 1.b et Fig. 1.c), puisqu'on relève que les classes obtenues par les relations de sous-catégorisation sont de moins bonne qualité (14 sur 23 classes sont des sous-classes de nos classes manuelles) que les classes obtenues à l'aide des relations de modification (20 sur 29 classes apparaissent valides).

Le type de relation – sous-catégorisation ou modification – semble donc avoir une incidence sur la qualité des associations produites avec la méthode distributionnelle lorsqu'on observe les proximités entre mots. Les relations adjectivales et apposition, plus lâches, permettent moins facilement de rendre compte du sens des noms. Les relations de sous-catégorisation paraissent plus adaptées pour cette tâche, mais la supériorité de l'analyse à l'aide des relations de sous-catégorisation n'apparaît cependant pas réelle si l'on observe les classes obtenues à l'aide des coefficients de proximité, probablement du fait d'un nombre de relations syntaxiques moins important pour ces distributions syntaxiques.

[6] Les relations incluant d'autres prépositions comme *sur* ou *dans* n'ont pas été retenues car elles mettent en jeu des relations de sous-catégorisation ou de modification selon le contexte. Le logiciel Syntex ne fait pas la différence entre ces deux types de relations.



(a) Ensemble des relations syntaxiques

(b) Relations de sous-catégorisation

(c) Relations de modification

Fig. 1 : Classification par voisinage à partir des coefficients de proximité (jaccard) entre mots

Le tableau 2 résume les résultats des méthodes employées.

	Ensemble des relations de dépendance	Relations de sous-catégorisation	Relations de modification
Nombre de paires dégagées	292	76	582
Qualité estimée pour les paires obtenues (avec la mesure jaccard)	60,5%	63%	49%
Précision des classes obtenues avec la classification par voisinage (calculée à partir du jaccard)	20/27 (74%)	14/23 (61%)	20/29 (69%)

Tableau 2 : Comparatif des méthodes employées

6 Conclusion

Les méthodes d'analyse distributionnelle automatique appliquées à notre champ lexical n'apparaissent qu'en partie concluantes. Les voisins obtenus à partir des distributions syntaxiques apparaissent valides à 60% si l'on tient compte de l'ensemble des relations syntaxiques. Nos résultats sont cependant pratiquement toujours meilleurs que ceux que Grefenstette (1996) obtient avec l'analyse syntaxique en comparant ses résultats à l'aune du thésaurus Roget. Nos classes sont cependant plus lâches.

La prise en compte des seules relations de sous-catégorisation augmente la précision (63%), mais le rappel est plus faible du fait du faible nombre de relations envisagées. Les résultats paraissent plus intéressants pour les classes obtenues par voisinage à l'aide du coefficient de proximité (jaccard), surtout si l'on prend en compte l'ensemble des relations syntaxiques (sans privilégier les relations de sous-catégorisation ou les relations de modification). Les classes obtenues confirment souvent la classification manuelle, tout en proposant des regroupements plus fins, probablement très utiles pour l'accès onomasiologique que nous envisageons pour notre application d'aide à la rédaction.

Deux types de traitement linguistique pourraient probablement améliorer les résultats. Tout d'abord, il serait souhaitable de normaliser les relations syntaxiques et les ramener à des relations plus sémantiques. Par exemple, il n'y a pas lieu de distinguer la relation entre l'adjectif épithète et le nom, et celle qui lie l'adjectif attribut et le nom. En outre, pour pallier le manque de données, il pourrait être utile de regrouper les relations par classes sémantiques, en utilisant la méthode distributionnelle de façon incrémentale. Enfin, il apparaît indispensable d'explorer d'autres mesures de similarité, comme la mesure prox, qui prend en compte la productivité de la relation syntaxique, ce qui n'est pas le cas de la mesure de jaccard.

Pour une application linguistique comme la nôtre, la méthode peut néanmoins apparaître utile, si les données obtenues sont validées manuellement. Le linguiste pourra ainsi partir des classifications obtenues automatiquement, observer les contextes partagés dans le corpus et corriger les données. Comme en terminologie, la méthode distributionnelle sera ainsi conçue comme une aide à la décision pour le lexicologue.

Remerciements

Tout d'abord, un très grand merci à Didier Bourigault qui m'a fourni les résultats de l'analyseur Syntex ainsi que les résultats du système d'analyse Upery et a relu une première version de ce papier. Merci également à Kjersti Fløttum, de l'Université de Bergen, qui m'a permis d'utiliser le corpus KIAP. Toute ma reconnaissance également à Christophe, le roi de Java, pour son aide. Merci aussi à Cécile Frérot pour ses conseils et à Olivier Kraif pour sa relecture d'une première version de ce papier.

Références

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. Actes de *la 9ème conférence annuelle sur le Traitement Automatique des Langues* (TALN 2002), Nancy, 2002, 75-84.

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P. & OZDOWSKA S. (2005), Syntex, analyseur syntaxique de corpus. Actes des *12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, France.

BOURIGAULT D., LAME G. (2002). Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit, in *TAL*, 43-1.

CRUSE D.A. (1986). *Lexical Semantics*. Cambridge, London : Cambridge University Press (Cambridge Textbooks in Linguistics).

GALY E., BOURIGAULT D. (à paraître). Analyse distributionnelle de corpus de langue générale et synonymie. *Actes JLC* 2005. Lorient.

GREFENSTETTE G. (1996). Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches. In Boguraev, B. and Pustejovsky, J., editors, *Corpus Processing for Lexical Acquisition*. Cambridge, Massachusset : MIT Press, 205-216.

GROSS G. (1994). Classes d'objets et description des verbes. *Langages* 115 , 15-30.

HABERT, B. AND ZWEIGENBAUM, P. (2003). Classer les mots : sémantique à gros grain et méthodologie harrissienne. *Revue de Sémantique et Pragmatique*, (12), 101–119.

HARRIS Z., GOTTFRIED M., RYCKMAN T. (1989). *The Form of Information in Science, Analysis of Immunology Sublanguage*. Kluwer Academic Publisher, Dordrecht, The Netherlands, 1989.

KRAIF O., TUTIN A. (2006). Des corpus bilingues alignés annotés sémantiquement pour l'aide à la rédaction: application aux collocations de la langue scientifique générale. *Aide à la rédaction - Apports du Traitement Automatique des Langues, Journée d'étude l'ATALA*, Paris.

NAZARENKO A., ZWEIGENBAUM P. , HABERT B, BOUAUD J. (2001). Corpus-based Extension of a Terminological Semantic Lexicon. *Recent Advances in Computational Terminology*. Amtserdam : John Benjamins, 327-351.

PECMAN M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique,* Thèse de doctorat, Université de Nice Sophia Antipolis, décembre 2004.

PHAL A. (1971). *Vocabulaire général d'orientation scientifique (V.G.O.S.) - Part du lexique commun dans l'expression scientifique*. Paris : Didier, Crédif.

Une expérience de compréhension en contexte de dialogue avec le système LOGUS, approche logique de la compréhension de la langue orale

Jeanne VILLANEAU
Valoria Université de Bretagne Sud,
Jeanne.Villaneau@univ-ubs.fr

Résumé. LOGUS est un système de compréhension de la langue orale dans le cadre d'un dialogue homme-machine finalisé. Il est la mise en œuvre d'une approche logique qui utilise différents formalismes afin d'obtenir un système robuste mais néanmoins relativement extensible. Cet article décrit essentiellement l'étape de compréhension en contexte de dialogue implémentée sur LOGUS, développée et testée à partir d'un corpus de réservation hôtelière enregistré et annoté lors des travaux du groupe MEDIA du projet technolangue. Il décrit également les différentes interrogations et conclusions que peut susciter une telle expérience et les résultats obtenus par le système dans la résolution des références. Concernant l'approche elle-même, cette expérience semble montrer que le formalisme adopté pour la représentation sémantique des énoncés est bien adapté à la compréhension en contexte.

Abstract. LOGUS is a spoken language understanding system usable in a man-machine dialogue. It is based on a logical approach where various formalisms are used, in order to achieve a robust but generic and extensible system. Implementation of a context-sensitive understanding is the main topic of this paper. Processing and tests were carried out from a hotel reservation corpus which was recorded and annotated as part of the work handled by the technolangue consortium'sMEDIA subgroup. This paper also describes the various questions raised and conclusions drawn from such an experiment, as well as the results achieved by the system for anaphora resolution. This experiment shows that the formalism used in order to represent the meaning of the utterances is relevant for anaphora resolution and in-context understanding.

Mots-clés : compréhension automatique de la parole, résolution des références, dialogue oral homme-machine.

Keywords: man-machine dialogue, spoken language understanding, anaphora resolution.

1 Introduction

Dans un système de dialogue oral Homme-Machine (DOHM), le module de « compréhension » de la langue orale spontanée remplit une tâche essentielle : à partir de la liste ou du graphe de mots que lui transmet le module de reconnaissance de la parole, il doit construire une structure sémantique qui puisse rendre compte du sens du message de l'utilisateur et qui soit utilisable par le module de dialogue.

Lorsque le système de DOHM est conçu pour une tâche très restreinte, horaires de train ou d'avion par exemple, cette interprétation du message peut se limiter à la détection d'une séquence de concepts, sur la base de structures sémantiques prédéfinies. Mais, lorsque le domaine d'application s'élargit, cette prédéfinition des requêtes devient plus complexe et la compréhension requiert d'autres approches (van Noord *et al.*, 1999).

LOGUS est un système de compréhension de la parole spontanée dans le cadre d'un DOHM conçu pour des domaines restreints, mais néanmoins plus étendus que les domaines où opèrent la plupart des systèmes actuellement opérationnels. Il correspond à la mise en œuvre d'une approche logique qui utilise différents formalismes pour combiner des outils syntaxiques et sémantiques. La prise en compte du contexte de dialogue est l'un des éléments essentiels de la « compréhension ». Particulièrement, l'approche utilisée pour la résolution des références reste symbolique et logique et vient ainsi en complément de celle utilisée dans la conception générale du système.

Cet article présente essentiellement les travaux d'implémentation de la compréhension en contexte de dialogue réalisés sur le système LOGUS à partir du corpus MEDIA. Après une brève exposition dans la section 2 des principes qui ont présidé à la conception du système LOGUS, la section 3 décrit le cadre du projet MEDIA et fait une brève analyse de son corpus, afin de dégager l'intérêt de son utilisation pour une telle expérimentation. Les principes de la compréhension en contexte et, plus précisément, de la résolution des références mises en œuvre dans LOGUS sont présentés dans la section 4. La section 5 présente une analyse quantitative et qualitative des résultats. L'article se termine avec la discussion et les conclusions présentées dans la section 6.

2 LOGUS : une utilisation de formalismes logiques pour la compréhension

LOGUS est un système de compréhension de l'oral spontané pour l'interrogation orale d'une base de données[1]. Il a été conçu pour fonctionner dans un domaine sensiblement plus large que ceux habituellement considérés pour ce type d'applications, où une représentation sémantique de l'énoncé par listes préconstruites d'attributs-valeurs s'avère suffisante. Néanmoins, l'analyse s'appuie sur une connaissance sémantique du domaine qui doit donc rester bien délimité et relativement étroit.

À partir d'une liste de mots issue d'un module de reconnaissance de la parole, LOGUS produit une formule logique qui représente le sens de l'énoncé. Le formalisme utilisé est adapté de la logique illocutoire de D. Vanderveken (Vanderveken, 2001) ; la formule logique s'obtient par application d'un acte de langage (sa *force propositionnelle*) à une structure construite à partir des « objets » de l'énoncé connus du système (son *contenu propositionnel*). La représentation sémantique peut également être représentée sous la forme d'un graphe conceptuel à la Sowa (Sowa, 2001). La figure 1 donne un exemple de la structure sémantique obtenue à partir d'un énoncé du corpus MEDIA (cf. 3.1).

L'analyse de l'énoncé est incrémentale et progressive ; elle se fait par étapes qui utilisent successivement différents formalismes logiques.

[1] Pour une description plus détaillée du système, on peut consulter les références suivantes : (Villaneau *et al.*, 2004; Villaneau, 2003).

L'énoncé :

« je souhaiterais réserver dans un hôtel Mercure trois étoiles à Belfort pour les quatre derniers jours de juin »

Sorties LOGUS :

```
(vouloir  (de   (reservation [(date (num_mois (derniers (entier 4)) (nom "juin")))]) |
                (hotel [    (marque_hotel (nom "Mercure")),
                            (etoiles (entier 3)),
                            (lieu (ville [(identification (nom "Belfort"))]))])
```

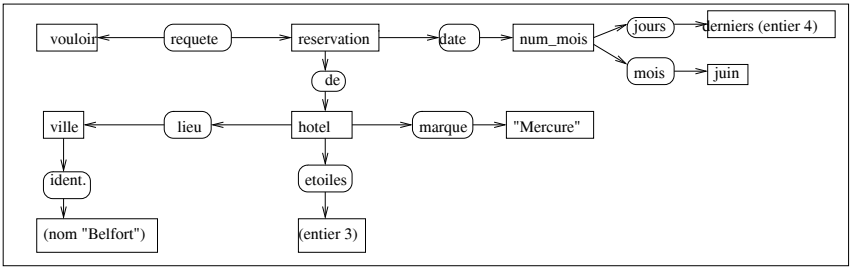


FIG. 1 – Un énoncé du corpus MEDIA et la sortie LOGUS correspondante

– Un lexique permet d'attacher à chaque mot « connu » une ou plusieurs « définitions ».
– Une première analyse partielle rattache les mots grammaticaux à leur tête lexicale. Cette étape peut être considérée comme un *« chunking minimaliste »*. Elle utilise des règles adaptées de celles des grammaires catégorielles de type AB et les termes simplement typés du λ-calcul (Villaneau & Antoine, 2004).
– Les étapes suivantes s'appuient sur une connaissance du domaine (ontologie) qui décrit le type des concepts du domaine d'application et les liens sémantiques qui peuvent les réunir. L'analyse de ces liens dans l'énoncé utilise les règles génériques d'une grammaire qui combinent les indices syntaxiques et sémantiques des différents composants. Ces règles sont appliquées en cascade, avec un assouplissement progressif des contraintes syntaxiques.
– La dernière étape prend en compte le contexte du dialogue pour compléter et préciser la compréhension de l'énoncé (cf. section 4). C'est essentiellement à l'expérimentation sur le corpus MEDIA de cette contextualisation qu'est consacré cet article.

3 Le projet MEDIA et son corpus

3.1 Le corpus MEDIA

Le projet MEDIA (programme Technolangue/Evalda) avait pour objectif l'évaluation de différents systèmes de compréhension en dialogue Homme-Machine, hors et en contexte de dialogue. Les partenaires du projet avaient choisi d'enregistrer un corpus à partir d'un serveur de réservation hôtelière. Le corpus enregistré par ELDA (appellé corpus MEDIA par la suite) pour la campagne d'évaluation comporte 1250 dialogues : les 250 utilisateurs du système ont interrogé le système suivant différents scenarii de réservation d'hôtels, élaborés par les partenaires

du projet. L'enregistrement s'est fait suivant le principe du Magicien d'Oz : les locuteurs ont dialogué avec un système simulé à leur insu par un opérateur humain. La figure 2 donne un extrait de dialogue du corpus. Dans cet exemple, *Ut* désigne l'utilisateur du système et *Co* le compère qui simule le système. Les expressions soulignées indiquent les marques linguistiques qui renvoient à une résolution des références en contexte de dialogue.

...	...
Ut	**"réserver une chambre simple du premier au six décembre dans trois endroits différents"**
Co	*"dans quelle ville souhaitez vous vous rendre"*
Ut	**"alors du un au trois décembre je veux être à Chalon Sur Saône du trois au cinq décembre à Bourg en Bresse et du cinq au six à Chaumont"**
Co	*"veuillez patienter nous recherchons cette information à Chalon Sur Saône du premier au trois décembre deux hôtels correspondent à votre demande le Saint Régis et le Ibis centre ville voulez vous des détails ou réserver dans l un de ces hôtels"*
Ut	**"hum des détails"**
Co	*"quelle information voulez vous"*
Ut	**"euh je voudrais le prix de la chambre"**
Co	*"'a l hôtel le Saint Régis la chambre individuelle est à cinquante euros 'a l hôtel Ibis centre ville la chambre individuelle est à cinquante euros souhaitez vous faire une réservation dans l un de ces hôtels"*
Ut	**"euh est ce que l un de ces hôtels accueille les animaux et est ce qu il y a un tennis"**
...	...
Co	*"... souhaitez vous réserver dans l un de ces hôtels"*
Ut	**"oui"**
Co	*"si oui lequel"*
Ut	**"euh le premier"**
...	...

FIG. 2 – Extrait de dialogue du corpus MEDIA

Le corpus ainsi enregistré a ensuite fait l'objet d'une transcription manuelle par ELDA, puis d'une annotation sémantique suivant les règles d'un manuel d'annotation mis au point par les partenaires du projet[2] : dans l'annotation sémantique hors-contexte, chaque énoncé est divisé en segments conceptuels « porteurs de sens ». À chacun d'entre eux est attribué un triplet *(mode, attribut, valeur)* ; des spécifieurs sont attachés aux attributs, afin de préciser les liens entre les différents concepts[3]. La figure 3 donne un énoncé extrait du corpus MEDIA avec son annotation sémantique.

Dans l'annotation sémantique en contexte de dialogue, les expressions référentielles portent les numéros des segments conceptuels auxquels elles renvoient.

[2]La mesure d'accord entre annotateurs (kappa) se situe au-dessus de 80%.

[3]Pour plus de détails, on peut consulter (Devillers *et al.*, 2004).

je souhaiterais réserver	+ :command-tache :reservation
dans un hôtel Mercure	+ :hotel-marque-reservation :mercure
trois étoiles	+ :hotel-etoile :3etoile
à Belfort	+ :localisation-ville-hotel :belfort
pour les quatre	+ :nombre-temps-reservation :4
derniers	+ :temps-axetps-reservation :dernier
jours	+ :temps-unite-reservation :jour
de juin	+ :temps-mois-reservation :6

FIG. 3 – L'énoncé de la figure 1 et son annotation MEDIA

3.2 Évaluation hors-contexte

Le système LOGUS a participé à la campagne d'évaluation hors-contexte (Bonneau-Maynard *et al.*, 2006). Il y a obtenu des résultats honorables mais sans grande signification[4]. En effet, la principale difficulté rencontrée pour la participation de LOGUS à cette campagne ne fut pas l'adaptation du système à la tâche, mais bien la transformation de la formule logique obtenue en la suite ordonnée de triplets *(mode, attribut, valeur)* demandée par MEDIA. Les sorties LOGUS sont globales : il y a « oubli » de l'ordre des mots et de la forme linguistique attachée à l'expression des requêtes. Comme on pouvait le craindre, l'annotation « collée au texte » de MEDIA s'est révélée souvent difficile voire impossible[5] à reconstituer à partir de la représentation sémantique finale : ainsi les deux tiers des erreurs relevées pour LOGUS ont été imputables à la transformation des sorties du système en la liste ordonnée des triplets demandée.

3.3 Les références dans Media

L'un des principes retenus par les partenaires MEDIA était que seules les expressions se rapportant à des références hors énoncé devaient être prises en considération.

Étant donné le domaine d'application retenu pour le projet, la résolution des références ne porte que sur quatre types d'objets : les hôtels, les chambres, les tarifs et les dates. Par ailleurs, les dialogues MEDIA sont généralement assez simples. Dans un dialogue standard du corpus, les énoncés les plus complexes sont ceux où l'utilisateur expose ses exigences. Ensuite, le compère pose des questions et, le plus souvent, l'utilisateur lui donne des réponses courtes et elliptiques. Malgré tout, les expressions anaphoriques y sont très diverses et représentatives de l'ensemble des difficultés classiquement rencontrées lors de la résolution des références.

Le corpus contient par exemple un grand nombre d'expressions définies et toutes les classes d'anaphores qui leur correspondent. Par exemples, suivant la classification proposée par C. Gardent et H. Manuélian (Gardent & Manuelian, 2005), *« les »* dans l'expression *« les animaux »* de l'extrait de dialogue donné figure 2 est une description autonome : elle ne donne pas lieu à une résolution. Dans ce même extrait le *« le »* de l'expression *« le prix de la chambre »* est une description associative ; comme son référent se trouve dans l'énoncé, il n'y a pas de résolution suivant les conventions MEDIA. En revanche, *« la »* dans cette même expression peut être

[4]Le système a été classé quatrième derrière les 2 systèmes du LIMSI et le système du LORIA (approche symbolique) et devant le système du LIA (approche stochastique).

[5]Sauf à remodeler profondément le système, solution a priori exclue.

considérée comme une description contextuelle, liée au référent des hôtels précités. Mais ce *« la »* est également coréférentiel dans la mesure où *« la chambre »* s'identifie avec la chambre demandée par l'utilisateur. Le choix MEDIA retient d'ailleurs les deux typologies puisque les segments référentiels de l'annotation contextuelle contiennent les caractéristiques de la chambre demandées par l'utilisateur (*chambre simple*) et les propriétés des hôtels proposés par le compère. Des expressions anaphoriques similaires sont introduites par des adjectifs démonstratifs : *cet hôtel, ces deux chambres, etc.*.

Le corpus contient également un très grand nombre d'expressions anaphoriques incluant une notion d'ordre : *le premier, le dernier, le deuxième, etc.* ou une notion d'exclusion : *l'autre, les autres, les deux autres, un autre, d'autres*.

On trouve également des pronoms qui désignent des référents au sens MEDIA : *je la réserve*, *est-ce qu'ils acceptent les chiens*, etc. alors que, par convention MEDIA, l'expression *« il y a »* n'est pas coréférentielle.

4 LOGUS : compréhension en contexte

4.1 Les principes généraux de la compréhension en contexte

Le principe général adopté pour la résolution des références dans LOGUS reste le même que celui qui prévaut à la compréhension hors contexte : combiner les critères syntaxiques et les critères sémantiques, ceux-ci prévalant sur ceux-là. En effet, si, dans les textes, les critères syntaxiques sont généralement plutôt bien respectés (Boudreau & Kittredge, 2005), il est loin d'en être de même à l'oral où, généralement, l'implicite domine. Le corpus MEDIA permet d'illustrer ces affirmations : par exemple, l'une des formulations les plus fréquentes dans le corpus MEDIA pour demander si un hôtel accepte les animaux est : *« est-ce qu'il acceptent les chiens »*. Cette utilisation du pluriel est si fréquente qu'il est difficile de penser qu'il s'agit là d'une faute de syntaxe. Elle correspond ici plutôt à une ellipse pour les *« gens de l'hôtel »*. On trouve également des expressions telles que *« celle à cinquante euros »* alors même que le référent logique est un hôtel. Il s'agit bien là encore d'une ellipse pour *« la chambre de l'hôtel »*.

L'une des relations sémantiques fondamentales utilisée pour la construction de la représentation du sens de l'énoncé dans LOGUS indique une dépendance entre deux objets. Cette relation conceptuelle générique, désignée par *« de »*, inclut par exemple les relations *partie-tout* ; elle permet de construire les « chaînes d'objets ». Ainsi *« le prix d'une chambre à l'hôtel Ibis »* correspond à la chaîne : *(tarif []) de (chambre []) de (hotel [(marque "Ibis")])* où l'objet « terminal » est *(hotel [(marque "Ibis")])*, en l'occurrence une entité nommée. La notion qui prévaut à la compréhension en contexte de dialogue pour le système LOGUS est celle de complétion des chaînes d'objets : dans un énoncé, une propriété ou un sous-objet peuvent être complétés par des chaînes de sur-objet du contexte, si cette complétion a un sens, donc si l'ontologie du domaine le permet. Par exemple, pour un énoncé *« quels sont les tarifs »* l'objet *tarif* serait automatiquement complété par la chaîne *(chambre []) de (hotel [(marque "Ibis")])* si cette chaîne est l'objet contextuel le plus proche.

4.2 LOGUS : mise en œuvre de la résolution des références pour le corpus MEDIA

Les dialogues du corpus MEDIA correspondent à un jeu de rôles relativement simple. Le locuteur énonce des contraintes ; le système propose des noms d'hôtels qui sont censés les satisfaire. Au cours du dialogue, l'utilisateur fait évoluer ses exigences et il y a succès si un accord finit par être trouvé entre celles-ci et les propositions du compère. Dans la pratique, les références liées au dialogue portent essentiellement sur les hôtels proposés par le compère et les sous-objets ou propriétés de ces hôtels liés aux exigences du demandeur.

L'objectif étant de pouvoir mesurer objectivement les performances du système à partir des exigences MEDIA, il convenait de respecter les principes généraux de l'annotation et de se limiter aux références qui correspondaient à des indices linguistiques explicites. Les principes généraux exposés dans le paragraphe précédent ont donc dû être largement amendés. Plus précisément, les références renvoyant à des chambres ou à des tarifs ont été traitées conformément à ces principes, avec un recopiage de la chaîne d'objets correspondante jusqu'au sur-objet de cette chaîne : l'hôtel concerné. Afin de faire un choix parmi les chaînes d'objets contextuels désignées comme sémantiquement possibles par l'ontologie, un traitement particulier a dû être appliqué pour chaque forme linguistique de la référence.

Chacun de ces traitements comporte généralement plusieurs niveaux : une première recherche où critères syntaxiques et sémantiques sont respectés, suivie d'un relâchement progressif des contraintes. Par exemple, *« le premier hôtel »* est d'abord recherché comme le premier élément de la dernière liste d'hôtels énoncés par le compère. Cependant, rien n'est moins sûr que cette liste existe. Les alea du dialogue, tours de parole interrompus par exemple, font que le compère n'a pas forcément proposé les différents hôtels dans un seul tour de parole : si donc cette première recherche ne rend pas de résultat, *« le premier hôtel »* sera alors recherché comme le premier hôtel proposé par le compère. De la même manière, chaque expression de *« autre »* : *les deux autres, un autre, l'autre,* etc. donne lieu à une stratégie de recherche particulière.

La résolution doit également prendre en compte les erreurs potentielles des locuteurs : erreurs de genre ou de nombre. Des exemples en ont déjà été donnés dans le paragraphe précédent (cf. 4.1) sous la forme d'ellipses ; il peut aussi s'agir parfois de véritables erreurs de la part du locuteur *« ces deux hôtels »* alors qu'il y a trois hôtels par exemple. Dans la résolution, ce type de contraintes sur les quantités exprimées ne sont relâchées qu'en tout dernier lieu. Il n'est d'ailleurs pas certain qu'elles devraient l'être dans un véritable système : il serait en effet sans doute plus pertinent que le système signifie à son interlocuteur qu'il ne l'a pas compris.

5 Analyse des résultats

Les tests ont été faits sur 100 dialogues pris au hasard dans le corpus annoté parmi ceux qui n'avaient pas servi au développement du système. Le tableau 1 donne les résultats chiffrés ainsi obtenus, et ce, de deux façons différentes. Les premiers sont calculés à partir des segments conceptuels définis dans l'annotation du corpus. Les seconds sont obtenus à partir des objets MEDIA eux-mêmes, un objet étant en général défini par plusieurs segments. Par exemple, un hôtel est en général référencé par deux segments conceptuels : son nom et sa ville. Une erreur sur l'un d'entre eux correspond en fait à une erreur sur l'objet lui-même. En revanche, on peut considérer que l'identification de la taille, de la date et de l'hôtel suffisent à référencer

une chambre, même si le (ou les) segments conceptuels qui précisent son prix a été oublié. Lorsque ces nombres ont été collectés, la différence entre les deux méthodes de calcul semblait flagrante. Or, si les résultats obtenus peuvent être très différents lorsqu'ils se rapportent à un ou deux dialogues, il est étonnant de constater que sur l'ensemble des dialogues testés, ils sont finalement globalement très comparables[6].

Nb de segments MEDIA (A)	Segments corrects (C)	Segments incorrects (I)	Rappel R=C/A)	Précision P=C/(C+I)	F 2RP/(R+P)
572	405	42	0,71	0,91	0,80
Nb d'objets MEDIA (A')	Objets corrects (C')	Objets incorrects (I')	Rappel R'=C'/A'	Précision P'=C'/(C'+I')	F' 2R'P'/(R'+P')
212	155	19	0,73	0,89	0,80

TAB. 1 – Résolution des références dans le corpus MEDIA : résultats chiffrés

Qualitativement, on peut classer les fautes faites par le système en quatre catégories.

- Comme l'indique le taux relativement bas du Rappel, la première cause d'erreurs est l'absence de détection de certaines références. Les articles définis sont particulièrement redoutables à cet égard. Par exemple, il n'est pas évident de savoir si une condition sur *« les chambres »* se rapportent ou non aux hôtels proposés précédemment. Par ailleurs, à l'instar du *« it »* de la langue anglaise (Boyd *et al.*, 2005), le pronom personnel *« il »* a beaucoup d'occurrences non référentielles et mériterait un traitement spécifique qui n'est actuellement pas réalisé.
- Certaines erreurs sont dues aux difficultés de compréhension des... énoncés du compère (cf. la discussion de la section suivante). Dans un énoncé tel que *« Astor Sofitel Novotel arc de triomphe Libertel Arc de triomphe »*, une segmentation correcte pour détecter les trois hôtels proposés n'est pas si évidente.
- D'autres erreurs sont dues à une mauvaise compréhension de la référence elle-même. Ainsi, pour des expressions telles que *« la deuxième proposition »*, *« celui qui reste »*, *« celui qui est près de l'autoroute »*, LOGUS donne une référence erronée.
- Il semble relativement facile de corriger une bonne partie des bugs précités. En revanche, pour résoudre certaines références, il faudrait munir le système d'une connaissance du contexte autrement plus complexe que celle dont il est actuellement pourvu. Par exemple, dans l'un des dialogues on a les tours de parole suivants :

  Compère : *« ... il reste cinq cents chambres disponibles »*

  Compère/Utilisateur : tours de parole concernant la date

  Utilisateur : *« oui vous me <u>la</u> réservez »*

  Le *« la »* fait référence aux cinq chambres demandées par l'utilisateur en début de dialogue, dans des tours de parole relativement éloignés. En relâchant les contraintes de nombre, LOGUS choisit les cinq cents chambres...

En conclusion, les traitements utilisés n'étant pas très sophistiqués, il serait sans doute relativement facile d'augmenter le rappel sans nuire à la précision. En revanche, un certain nombre de références font appel au « sens commun », celui qui est si difficile à cerner et à implémenter...

[6] Aucune comparaison directe n'est malheureusement possible avec les deux participants de la campagne d'évaluation en contexte de MEDIA puisque la façon de procéder pour l'évaluation de LOGUS a éliminé les contraintes dues à la forme requise par MEDIA et les erreurs qu'elle implique.

6 Discussion et conclusion

Une première critique possible de cette expérience est le caractère quelque peu artificiel de l'exercice.

– La compréhension en contexte de dialogue sur ce corpus a demandé que soit implémentée une compréhension des énoncés compère. Cette tâche n'aurait pas lieu d'être dans le module de compréhension d'un véritable système. Ceci dit, comprendre un énoncé-système est beaucoup plus simple que comprendre un énoncé-utilisateur puisque les formes linguistiques utilisées sont connues et stéréotypées. Mais tout aussi consciencieux et appliqués que soient les compères qui simulent un système dans un corpus élaboré par la technique du Magicien d'Oz, ils ne peuvent pas simuler parfaitement un véritable système. Les expressions qu'ils utilisent ne sont pas entièrement stéréotypées et laissent place à une assez large variabilité. Par ailleurs, il leur arrive également de se tromper, c'est à dire, en l'occurrence, de proposer des réponses non conformes à celles que pourrait proposer le système.
– On peut également discuter le bien-fondé du choix fait dans MEDIA d'avoir utilisé un corpus dont la retranscription « gomme » les erreurs dues à la reconnaissance de la parole car on sait qu'il s'agit là d'une des plus grandes difficultés rencontrées par les systèmes de compréhension de la langue orale. On peut aussi défendre la position selon laquelle les problèmes traités sont suffisamment complexes pour mériter d'être clairement séparés.

L'utilisation du corpus MEDIA pour tester le système LOGUS présente un autre type d'inconvénients, liés à la nature même du système testé.

– La résolution des références faite dans MEDIA ne correspond pas vraiment à l'approche de la compréhension en contexte envisagée pour LOGUS. Dans MEDIA, la résolution des références est entièrement basée sur l'existence d'indices linguistiques ; par exemple, l'expression *« est-ce qu'ils acceptent les chiens »* demande la résolution d'une référence à cause du pronom *ils* alors que la question posée sous la forme *« est-ce que les chiens sont acceptés »* n'est pas référentielle. Dans l'approche LOGUS, les deux expressions appellent une résolution identique. L'acceptation des animaux est une propriété relative à un hôtel : la compréhension en contexte demande que soit recherché le (ou les) hôtels éventuellement concerné(s) par cette interrogation. Tester les principes de la compréhension en contexte de LOGUS à partir du corpus MEDIA a donc demandé que soient mis de côté certains des principes fondamentaux de la compréhension contextuelle.
– Enfin, comme il a été dit précédemment, le système LOGUS a été conçu pour essayer d'élargir les domaines potentiels de la compréhension en dialogue homme-machine. Un domaine tel que la réservation hôtelière reste trop étroit pour valider complètement l'approche utilisée. La représentation sémantique des énoncés choisie par les partenaires MEDIA en triplets *(mode, attribut, valeur)* reste pertinente pour une telle application et les formules logiques construites par LOGUS peuvent apparaître comme inutilement complexes.

En même temps et malgré les réserves précédentes, le corpus MEDIA est composé de véritables dialogues dans lesquels, malgré leur relative simplicité, on retrouve la plupart des problèmes classiques liés à la résolution des références. L'annotation des références en fait un corpus de langue française parlée très intéressant pour la mise au point de systèmes de compréhension dans le domaine du DOHM.

Les résultats obtenus par LOGUS au cours de cette expérience sont encourageants : ils semblent en effet prouver que la notion de chaînes d'objets utilisée pour la représentation sémantique est un point de départ solide pour la résolution des références et la compréhension en contexte.

Remerciements

Merci à tous les membres du consortium MEDIA pour toutes les discussions que nous avons pu avoir et sans lesquelles ce travail n'aurait pas été possible.

Références

BONNEAU-MAYNARD H., AYACHE C., BECHET F., A. A. D., KHUN A., LEFEVRE F., MOSTEFA D., QUIGNARD M., ROSSET S., SERVAN C. & VILLANEAU J. (2006). Results of the french evalda-media evaluation campaign for literal understanding. In *the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, p. 2054–2059, Genoa, Italy.

BOUDREAU S. & KITTREDGE R. (2005). Résolution des anaphores et détermination des chaînes de coréférences. *Traitement Automatique des Langues (TAL)*, **46**(1), 41–69.

BOYD A., GEGG-HARRISON W. & BYRON D. (2005). Identifying non-referential *it* : a machine learning approach incorporating linguistically motivated patterns. *Traitement Automatique des Langues (TAL)*, **46**(1), 71–90.

DEVILLERS L., BONNEAU-MAYNARD H., ROSSET S., PAROUBEK P., MOSTEFA D., CHOUKRI K., CHARNAY L., BOUSQUET C., VIGOUROUX N., BECHET F., ROMARY L., ANTOINE J.-Y., VILLANEAU J., VERGNES M. & GOULIAN J. (2004). The french evalda-media project : the evaluation of the understanding capabilities of spoken language dialogue systems. In *the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, p. 2131–2134, Lisboa, Portugal.

GARDENT C. & MANUELIAN H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues (TAL)*, **46**(1), 115–139.

SOWA J. (2001). Conceptuals Graphs. http ://users.bestweb.net/ sowa/cg/cgstand.htm.

VAN NOORD G., BOUMA G., KOELING R. & NEDERHOF M. (1999). Robust grammatical analysis for spoken dialogue systems. *Natural Language Engineering.*, **5**(1), 45–93.

VANDERVEKEN D. (2001). Universal Grammar and Speech Act Theory. In D. VANDERVEKEN & S. KUBO, Eds., *Essays in Speech Act Theory*, chapter 2, p. 25–62. Amsterdam, Philadelphia : John Benjamin.

VILLANEAU J. (2003). *Contribution au traitement syntaxico-pragmatique de la langue naturelle parlée : approche logique pour la compréhension de la parole*. PhD thesis, Université de Bretagne Sud, Vannes, France.

VILLANEAU J. & ANTOINE J.-Y. (2004). Categorials grammars used to partial parsing of spoken language. In *Actes de CG2004*, p. 244–258, Montpellier, France.

VILLANEAU J., RIDOUX O. & ANTOINE J.-Y. (2004). LOGUS : compréhension de l'oral spontané. *Revue d'intelligence artificielle*, **18**(5–6), 709–742.

Évaluation des performances d'un modèle de langage stochastique pour la compréhension de la parole arabe spontanée

Anis ZOUAGHI[1], Mounir ZRIGUI[1], Mohamed BEN AHMED[2]

[1] Labo RIADI (Unité de Monastir)
Université de Monastir, Faculté des sciences de Monastir
[2] Labo RIADI – Université de la Mannouba,
École nationale des sciences de l'informatique
Anis.Zouaghi@riadi.rnu.tn, Mounir.Zrigui@fsm.rnu.tn,
Mohamed.Benahmed@riadi.rnu.tn

Résumé. Les modèles de Markov cachés (HMM : Hidden Markov Models) (Baum et al., 1970), sont très utilisés en reconnaissance de la parole et depuis quelques années en compréhension de la parole spontanée latine telle que le français ou l'anglais. Dans cet article, nous proposons d'utiliser et d'évaluer la performance de ce type de modèle pour l'interprétation sémantique de la parole arabe spontanée. Les résultats obtenus sont satisfaisants, nous avons atteint un taux d'erreur de l'ordre de 9,9% en employant un HMM à un seul niveau, avec des probabilités tri_grammes de transitions.

Abstract. The HMM (Hidden Markov Models) (Baum et al., 1970), are frequently used in speech recognition and in the comprehension of foreign spontaneous speech such us the french or the english. In this article, we propose using and evaluating the performance of this model type for the semantic interpretation of the spontaneous arabic speech. The obtained results are satisfying; we have achieved an error score equal to 9.9%, by using HMM with tri-grams probabilities transitions.

Mots-clefs : analyse sémantique, modèle de langage stochastique, contexte pertinent, information mutuelle moyenne, parole arabe spontanée.

Keywords: semantic analysis, stochastic language model, pertinent context, overage mutual information, spontaneous arabic speech.

1 Introduction

On distingue deux grands courants d'approches pour la compréhension de la parole : les approches symboliques linguistiques (ou par règles), et les approches stochastiques. Le premier type d'approches se base sur une représentation préalable de la grammaire. Pour décrire cette grammaire, on utilise généralement l'un des formalismes existants tels que : HPSG, les grammaires lexicales fonctionnelles (LFG), etc. Quand au deuxième type

d'approche, les règles sont déduites directement à partir d'un corpus d'apprentissage. Depuis quelques années, la tendance est vers l'utilisation des modèles de langages stochastiques dans le domaine de la compréhension de la parole spontanée (Schwartz et al., 1996), (Minker, 1999), (Bousquet, 2002), etc. Cette tendance s'explique par le faite que les approches stochastiques offrent une alternative efficace aux approches par règles, concernant le coût global de développement du modèle, et la portabilité vers d'autres domaines. De plus, du fait que le locuteur parle d'une manière spontanée, les fautes de syntaxe ou de grammaire sont beaucoup plus fréquentes à l'oral qu'à l'écrit. C'est pour cela, qu'une analyse portant uniquement sur la syntaxe n'est souvent pas efficace. Ainsi, certains proposent pour faire face à ce problème, une analyse plus fine des phénomènes linguistiques de l'oral tels que (Van Noord et al., 1999) et (Antoine et al., 2003), ou une combinaison d'une analyse syntaxique et sémantique tels que (Villaneau et al., 2001), (Seneff, 1992), etc. Contrairement à la langue latine, la compréhension automatique de la parole arabe spontanée reste encore très peu abordée au niveau de la recherche scientifique. Durant les deux dernières décennies les efforts ont été plutôt concentrés sur la réalisation des analyseurs morphologiques et syntaxiques pour l'arabe tel que (Ouersighni, 2001). Malgré l'importance de la représentation et de l'analyse sémantique pour la réalisation de n'importe quel système de compréhension, il n'existe que quelques travaux qui s'intéressent à ce domaine en vue du traitement automatique de la langue arabe écrite et non pas parlée tels que (Haddad et al., 2005), (Meftouh et al., 2001), etc. Dans cet article, nous présentons le modèle de langage stochastique employé pour l'analyse sémantique de la parole arabe spontanée dans le cadre d'une application finalisée, ainsi que les résultats d'évaluation obtenus.

2 L'application finalisée considérée

2.1 Le domaine de l'application

Pour tester et estimer les paramètres du modèle de langage stochastique, nous avons utilisé un corpus représentant le domaine des renseignements ferroviaires. La principale raison de ce choix est la taille statistiquement représentative du corpus d'apprentissage dont nous disposons (voir tables 1 et 2). Ce corpus a été collecté en demandant à cent personnes différentes de formuler des énoncés relatifs aux renseignements ferroviaires. Donc c'est un corpus simulé et non pas réel.

Domaine	Taille (Mo)	Nombre d'énoncés	Nombre de mots	Nombre de locuteurs
Renseignements ferroviaires	*3,4*	*10000*	*85900*	*1000*

Table 1 : Caractéristiques du corpus de point de vue volume.

Nature de la tâche	Renseignements sur les:				Réservations	autres
Taux de sa représentation	horaires	trajets	tarifs	durées	*10,41 %*	*40,64%*
	28,7 %	*9,37 %*	*16,66 %*	*3,12 %*		

Table 2 : Caractéristiques du corpus de point de vue contenu.

2.2 Le corpus d'apprentissage

Le modèle de langage va servir à attribuer à chaque mot de l'énoncé transcrit par le module de reconnaissance de la parole un couple de traits sémantiques noté TS. Chaque couple TS est constitué de deux traits élémentaires : TS = (classe sémantique TSC, trait micro sémantique TSM). Le premier trait sert à déterminer la classe sémantique à laquelle appartient le mot à interpréter. Par exemple, toutes les villes du réseau ferroviaire sont représentées par la classe sémantique مدينة "medina "(ville). Pour l'application considérée, nous avons utilisé en tout 12 classes sémantiques différentes (voir table 3 ci-dessous).

Classes sémantiques TCS	Exemples d'instanciations
طلب (demande)	متى (quand)- بكم (combien)- أحب (je veux)- يوجد (existe)- etc.
حركة (mouvement)	الذاهب (qui va)- اذهب (je vais)- يصل (arrive)- etc.
مؤشر_حركة (Indice_mouvement)	إلى (à)- نحو (vers)- عبر (à travers)- من (de)- etc.
مؤشر_توقيت (Indice_horaire)	بتاريخ (à la date)- الساعة (l'heure)- etc.
رمز (référence)	هذا (ce)- هاته (cette)- etc.
مدينة (ville)	سوسة (sousse)- تونس (Tunis)- etc.
ربط (liason)	و (et)- etc.
عدد_تذاكر (nombre_billets)	تذكرة (biellet)- مكان (place)– تذكرتين (deux billets)– etc.
حس (bruit)	أن (que)- نهاركم (journée)- etc.
نوع_التذكرة (type_billet)	مسترس - ل- ذهاب_ و_إياب - للصغار - للطلبة - etc.
شرط (condition)	لا_ تتجاوز_أعمارهم (qui_ne_dépassent_pas_l'age)- etc.
عدد (nombre)	2 – 1 etc.

Table 3 : Les classes sémantiques considérées.

La méthode d'identification ou d'extraction de ces classes est présentée dans le paragraphe suivant (*2.3*). En ce qui concerne le deuxième trait du couple TS, c'est un trait micro sémantique qui permet de différencier le sens des mots appartenant à une même classe sémantique. Par exemple, ce trait permet de distinguer une ville de départ d'une ville de destination dans un énoncé donné. Nous signalons que les mots synonymiques ou possédant un même rôle sémantique possèdent le même couple de traits TS. Le nombre total des traits micro sémantiques TMS utilisés est 20 traits, soit presque le double des TCS. Ces traits sont les suivants : طلب _ توقيت (demande_prix) – طلب _ ثمن (demande_générale) - عام _ لحظة (correspondance) – عبور (départ) – انطلاق (destination) – وجهة (demande_horaire) – درجة (classe) – etc. Ainsi, pour estimer (moment) – ساعة (heure) – تاريخ (date) – يوم (jour) – les paramètres du modèle de langage stochastique, nous avons crée un corpus d'apprentissage (voir figure 1). Ce corpus a été obtenu en étiquetant au début manuellement une quantité (500) des énoncés du corpus collecté par un expert humain. Le principe d'étiquetage est d'attribuer à chaque mot significatif pour l'application un couple TS tel que défini ci haut. Les mots non significatifs ou vides sont éliminés lors de la phase du prétraitement du corpus

initial, et certains mots sont regroupés en une seule entrée. L'élimination des mots vides nous a permis de simplifier la complexité et réduire la taille du modèle. Ensuite, nous avons appliqué ce modèle pour l'étiquetage sémantique des 9000 énoncés restants, et ce par groupes de 500. Entre chaque étape d'étiquetage automatique, nous avons procédé à une vérification des résultats obtenus et une correction des paramètres a été établie chaque fois qu'il y a une détection d'erreurs. Enfin, les 500 énoncés restants nous ont servi pour l'évaluation de la performance du modèle. Ainsi, 95% du corpus a été consacré à l'apprentissage et 5% aux tests.

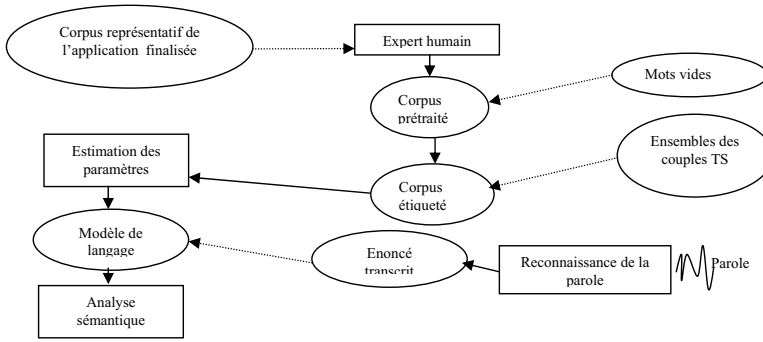


Figure 1 : Principe de l'estimation des paramètres du modèle de langage.

Les flèches en pointillés dans la figure 1, correspondent aux informations qui dépendent du domaine de l'application à modéliser.

2.3 L'extraction des classes sémantiques

Pour extraire les classes sémantiques de l'application, nous avons appliqué l'algorithme des K-means proposé par (McQueen, 1967), en utilisant l'information mutuelle moyenne IMm de (Rosenfeld, 1994) au lieu de la distance euclidienne pour mesurer la distance sémantique entre les différents mots du vocabulaire de l'application finalisée. Ceci, nous a amené à remplacer dans l'algorithme le critère d'évaluation $arg\ min_{j=1,\ldots,k}\ d^2(m_i, cgj)$ par $arg\ max_{j=1,\ldots,k}\ d(m_i, cgj)$ (voir figure 2). A part que cet algorithme, permet de faciliter la tâche d'identification des classes sémantiques, il a l'avantage d'être :

- Rapide face à des données de taille importante, puisqu'il converge à une vitesse linéaire de l'ordre de O(n.k.t) ; où n, k et t désignent respectivement le nombre des mots à classer, le nombre des classes sémantiques et le nombre d'itérations maximales.
- Et simple à implémenter.

Présentation de l'algorithme des k-means : -
Choisir d'une manière arbitraire les centres de gravité (cg1, cg2, cg3, …, cgk) des k classes sémantiques (cs1, cs2, cs3, …, csk).
Début
- *Etiquette* :
  Pour tout mot mi de m_1 à m_n faire
    Chercher la classe csk du mot mi en question :
      csk = arg $max_{j=1,\ldots,k}$ d(m_i, cgj) ;

où, d (m_i, cgj) = IMm(mi, cgj) = P(mi, cgj)× Log [P(mi / cgj) / P(mi).P(cgj)] + P($\overline{mi}$, $\overline{cgj}$)× Log [P($\overline{mi}$ / $\overline{cgj}$) / P($\overline{mi}$).P($\overline{cgj}$)] + P(mi, $\overline{cgj}$)× Log [P(mi / $\overline{cgj}$) / P(mi).P($\overline{cgj}$)] + P($\overline{mi}$,

cgj)× Log [P(mi / cgj) / P(mi).P(cgj)]
Recalculer le centre de gravité de la classe csk :
cgk = $1/N_k \sum_{mi \in csk}$ mi ; où N_k désigne dans cet algorithme le nombre de mots dans la classe csk.
Fin Pour.
- Arrêt du traitement si les centres de gravité sont inchangés.
- Retourner à *Etiquette* sinon.
Fin

Figure 2 : l'algorithme des k-means en utilisant l'IMm comme métrique.

Cependant, le problème principal de cette méthode est la dépendance du résultat du classement final des informations données en entrée (les k centres de gravité des k classes sémantiques à déterminer sont choisis d'une manière totalement arbitraire). Cette limite ne pose pas de problèmes pour nous, puisque nous avons utilisé cette méthode rien que pour aider et donner une idée à l'utilisateur (surtout si cet utilisateur n'est pas un expert du domaine) sur la classification possible des mots de l'application d'un point de vue sémantique. Cependant les cartes auto organisatrices de (kohonen, 1989) offrent une alternative efficace, pour ceux qui cherchent des meilleurs résultats de partitionnement (Jamoussi, 2004).

3 Modélisation stochastique

3.1 Description du système de compréhension

Le système de compréhension conçu permet de construire la représentation sémantique d'un énoncé, sous la forme d'un ensemble d'associations attributs/valeurs (ou formulaire), comme le montre l'exemple suivant : Enoncé transcrit : أريد حجز مكان بالقطار الذاهب إلى تونس. ″ouridou hajza makan bilqitar athaheb ila tuwnis″ → Je veux réserver une place dans le train allant à Tunis.

Représentation sémantique :

نوعية (Type) = طلب حجز (demande de réservation)
مدينة_انطلاق (ville_ départ) = &Villecourante
يوم_انطلاق (jour_ départ) = ?
ساعة_انطلاق (heure_ départ) = ?
مدينة_وجهة (ville_ destination) = تونس (Tunis)
عدد_مقاعد (nombre_places) = 1

La figure 3 ci-dessous, présente l'architecture générale du système de compréhension. On remarque bien que la déduction du sens d'un énoncé par ce système est le résultat de l'accomplissement des traitements successifs suivants :
- La segmentation de l'énoncé transcrit par le module de reconnaissance de la parole : ce traitement permet d'identifier les mots ainsi que les différentes phrases du message du locuteur. Un même message peut être constitué d'un ou plusieurs requêtes à la fois. D'où, il est nécessaire que le système puisse identifier les différentes requêtes du message, afin qu'il puisse interpréter la demande de l'utilisateur dans toute son intégralité.
- Le prétraitement de l'énoncé : ce prétraitement consiste comme pour le prétraitement du corpus collecté à éliminer par exemple les mots vides, à regrouper certains mots en une seule entrée, etc. Ce modèle permet de simplifier la complexité de la tâche de compréhension.
- Le décodage sémantique de l'énoncé : c'est-à-dire l'étiquetage de chaque mot de l'énoncé prétraité avec les couples TS correspondants.
- La construction du sens de l'énoncé, cette étape correspond à la phase de génération de

l'ensemble des paires attribut/valeur (ou formulaire).
Le décodage sémantique des énoncés prétraités repose sur un modèle de langage stochastique qui permet d'encoder les règles de la grammaire (*voir paragraphe suivant*) et sur un lexique sémantique décrit dans un fichier et contient tous les mots du vocabulaire de l'application. Ce lexique est un ensemble d'associations de la forme : Mot M / TS décrivant le sens du mot + P(W / TSC, TSM) qui est la probabilité d'utilisation de TS = (TSC, TSM) pour la description du sens du mot M.

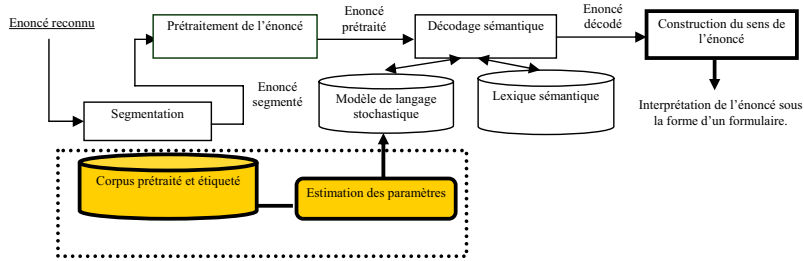


Figure 3 : Architecture du système de compréhension

3.2 Le modèle de langage

3.2.1 Le principe du décodage

Le modèle de langage que nous présentons ici, permet d'attribuer à chaque mot significatif un couple TS permettant de décrire sons sens. Comme nous l'avons signalé auparavant, nous avons choisi de représenter ce modèle à l'aide d'un modèle de Markov caché. Le principe du décodage sémantique est le suivant :
Nous considérons un énoncé constitué d'une suite de n mots : W = w1 w2 … wn
Cette suite de n mots est réduite à une suite de m mots après la phase du prétraitement de l'énoncé (élimination et regroupement de certains mots), où m ≤ n : W = w1 w2 … wm
Supposons que cette suite a été décodée via la suite de m couples de traits sémantiques suivante:TS=TS1TS2…TSm, ou encore TS= (TSC1,TSM1)(TSC2,TSM2)…(TSCm, TSMm).
Le but est alors de trouver les meilleures suites TS' connaissant W. Cette probabilité est calculée grâce au critère du maximum a posteriori : $P(TS' / W') = \text{Max}_{TS} P(TS / W) = \text{Max}_{TSC \times TSM} P(TSC, TSM / W)$
Ce qui donne en appliquant la formule de Bayes : $P(TSC, TSM / W) = P(W / TSC, TSM) \times P(TSC, TSM) / P(W)$
Nous avons ensuite utilisé l'algorithme de Viterbi (Rabiner et al., 1986), pour réaliser ce décodage.

3.2.2 La topologie du modèle

Nous avons considéré un modèle de Markov caché (HMM) à un seul niveau pour réaliser notre décodeur (voir figure 4). Chaque état du modèle markovien représente un couple TS et

les probabilités de transitions représentent les probabilités de passage d'un TS vers un autre. L'interprétation d'un mot dépend du contexte de l'énoncé, c'est-à-dire des relations de dépendances qu'il entretient avec les autres mots de l'énoncé. Comme ll montre la figure 4 suivante, nous avons considéré un HMM avec des probabilités tri-grammes de transitions entre les couples de traits sémantiques TSi des mots. Ce modèle contribue ainsi à la prédiction d'un couple de traits sémantiques TSi à partir des deux couples précédents TSi-1 et TSi-2.

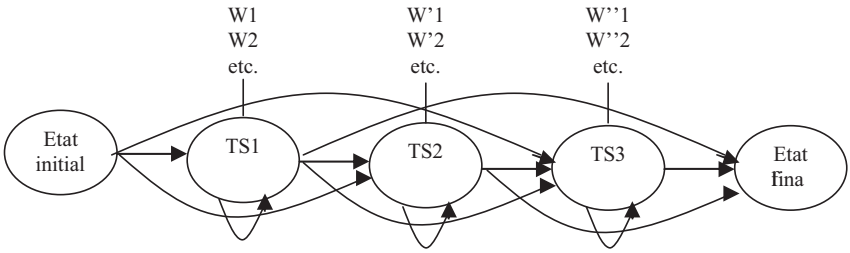


Figure 4 : Exemple de modélisation à l'aide d'un modèle de Markov caché à un niveau avec des probabilités tri-grammes de transitions entre les TSi

La réalisation de ce modèle nécessite principalement deux types d'informations :
- La manière d'agencement des couples TSi entre eux, sous la forme de probabilités tri-grammes de transitions entre les TSi :
P(TSi / TSi-1, TSi-2) = N (TSi, TSi-1, TSi-2) / N (TSi-1, TSi-2) ; où N (TSi, TSi-1,TSi-2) (resp. N (TSi-1, TSi-2)) est le nombre d'occurrence de TSi, TSi-1 et TSi-2 (resp. TSi-1 et TSi-2) ensemble.
- Et la probabilité d'émission de chaque mot du vocabulaire de l'application par chacun des couples TS définis. Un mot peut être décrit sémantiquement par plusieurs couples TS.
P (W/ TS) = N (W, TS) / N (TS) ; où N (W, TS) est le nombre fois de description de W par TS et N (TS) est le nombre total d'utilisation de TS.

3.2.3 Amélioration du modèle

En remarquant que ce n'est pas obligatoirement les mots précédant immédiatement le mot à interpréter qui ont une influence sémantique sur ce dernier, nous avons décidé d'employer lors de la phase de décodage du sens d'un mot que les TS des deux mots possédant la plus grande affinité sémantique avec celui-ci. Pour atteindre cet objectif, nous nous sommes basés sur la notion d'information mutuelle moyenne (Rosenfeld, 1994) qui permet de calculer le degré de corrélation ou de co-occurrence de deux mots donnés. Cette méthode nous a permis de ne plus utiliser systématiquement les TS des deux mots qui précédent immédiatement le mot à décoder.

4 Application du modèle et résultats

Pour tester la performance du modèle stochastique défini, nous avons utilisé les 500 énoncés du corpus collecté qui n'ont pas été employés lors de la phase d'estimation des paramètres du modèle de langage stochastique (*voir paragraphes 2.1 et 2.2*). Nous avons utilisé comme mesures de performances : - -

Le nombre total de mauvaises interprétations N_f défini comme suit : $N_f = N_{CS} + N_{MS}$, où N_{CS} et N_{MS} sont respectivement le nombre de TSC et le nombre de TSM incorrectement attribués par le système aux mots de l'énoncé.

- Le taux d'erreur du décodage sémantique : $Taux_{erreur} = N_f / N$; où N est le nombre total de traits TSC et TSM attribués à l'énoncé à interpréter.
- Le taux de précision est : $Taux_{précision} = N_C / N$; où N_C est le nombre des traits TSC et TSM correctement attribués.

La figure 5 suivante, présente les taux d'erreur et de précision trouvés. Ces taux sont répartis selon le type de renseignement demandé par l'utilisateur : demande de réservation (DR), ou de renseignements sur le trajet (DT), l'horaire (DH), le prix (DP), ou la durée du voyage (DD). On peut toujours aussi relever le taux d'erreurs des énoncés incorrectement décodés sémantiquement, en considérant le rapport entre les énoncés mal interprétés et le nombres total d'énoncés considérés dans le test (ici 500).

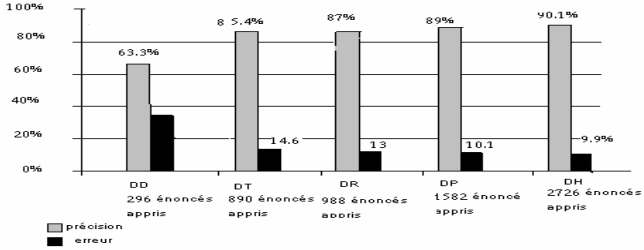


Figure 5 : Taux d'erreur et de précision selon le type de la demande de l'utilisateur et le nombre d'énoncés appris.

Le taux d'erreur réellement trouvé lors de la mesure de la performance de notre système est de l'ordre de 21,1%. En analysant davantage les résultats, nous avons conclu qu'un mauvais décodage est obtenu chaque fois qu'il y a un manque de données d'apprentissage. La figure 5 ci-dessus illustre bien ceci. En effet, d'après cette figure, nous remarquons que les résultats de décodage sont bons dans presque tous les types de renseignements demandés par l'utilisateur (DT, DR, DP et DH). Le plus mauvais décodage correspond aux énoncés de type DD. Ceci est dû au fait, que le nombre des énoncés DD considérés (3,12% du corpus) lors de la phase d'apprentissage du modèle de langage est insuffisant. En effet, nous avons constaté que seulement à partir de 1000 énoncés appris que notre système devienne performant. A partir de ce seuil, le taux d'erreur est inférieur à 11%. Au dessous de la barre de 500 énoncés, les résultats deviennent inacceptables. Le taux d'erreurs atteint 36,7% pour 296 énoncés appris, alors qu'il se restreint à 9,9% pour 2726 énoncés appris (voir figure 5). Donc, une mauvaise interprétation par notre système est due essentiellement à un manque de données d'apprentissage, et non pas au type ou à la topologie du modèle de langage utilisé. Nous avons aussi comparé ce modèle de langage employé par rapport à un modèle de langage avec des probabilités bi-grammes de transitions entre les TSi (1) et un modèle de langage avec des probabilités tri-grammes de transitions sans amélioration (2) (c-à-d sans considération des TS

des 2 mots influant sémantiquement sur le mot à interpréter). Nous avons trouvé que modèle (1) est efficace seulement lorsque le corpus d'apprentissage n'est pas assez volumineux (voir figure 6). En effet, plus l'ordre n d'un modèle n-grammes est petit, moins on a besoin de données d'apprentissage. Donc le modèle (1) peut être une alternative efficace au modèle utilisé (avec tri-grammes de transitions amélioré), dans le cas où on ne dispose pas de corpus assez volumineux représentatif du domaine de l'application à modéliser. Mais nous avons constaté que dés qu'il y a occurrence d'hésitations ou de mots inconnus précédant le mot à interpréter les modèles (1) et (2) deviennent aussi inefficaces.

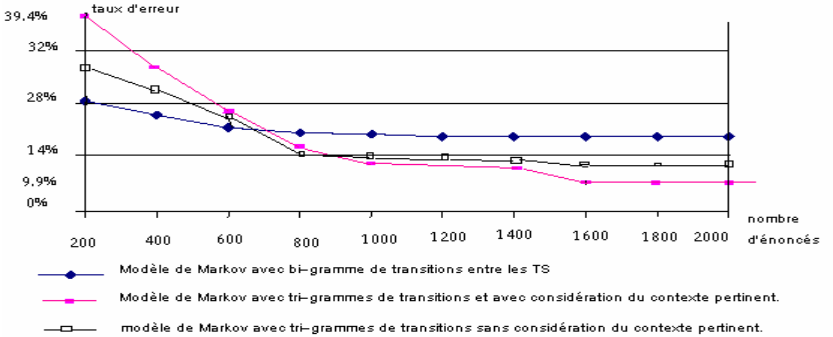


Figure 6 : Résultats de décodage obtenus en utilisant des modèles de Markov avec bi-grammes et tri-grammes de transitions avec et sans considération du contexte pertinent.

5 Conclusion

Nous avons présenté dans cet article le modèle de langage stochastique que nous avons employé pour le décodage sémantique de la parole arabe spontanée. Pour cela, nous avons utilisé un modèle de Markov caché à un seul niveau, avec des probabilités tri-grammes de transitions entre les couples de traits sémantiques TS. L'évaluation du modèle, en l'appliquant dans le domaine des renseignements ferroviaires a montré son efficacité. Nous avons atteint un taux de précision de l'ordre de 90,1% avec 2726 énoncés appris de type demandes d'horaires. Nous avons montré qu'en cas de manque de données d'apprentissage, un modèle de Markov caché à un seul niveau, avec des probabilités bi-grammes de transitions entre les TS est plus puissant. Ceci est vrai malheureusement que dans le cas d'énoncés non spontanés, c'est-à-dire ne contenant ni des hésitations ni des mots inconnus. Pour identifier les couples TS à employer pour l'interprétation des mots de l'énoncé, nous avons employé l'information mutuelle moyenne IMm de (Rosenfeld, 1994). Pour faciliter la tâche d'extraction des traits TSC d'une application, nous avons utilisé l'algorithme de partitionnement des K-means proposé par (McQueen, 1967). Cependant comme nous l'avons déjà signalé, nous avons utilisé cette méthode rien que pour aider et donner une idée à l'utilisateur sur la classification possible des mots de l'application d'un point de vue sémantique. Cependant les cartes auto organisatrices de (kohonen, 1989) offrent une alternative efficace, pour ceux qui cherchent des meilleurs résultats de partitionnement (Jamoussi, 2004).

Références

ANTOINE J-Y., GOULIAN J., VILLANEAU J. (2003), Quand le TAL robuste s'attaque au langage parlé: analyse incrémentale pour la compréhension de la parole spontanée, Actes de *TALN*.

Baum L.E., Petrie T., Soules G., Weiss N. (1970), A maximisation technique occurring in statistical analysis of probabilistic functions in Markov chains, *The Annals of Mathematical Statistics*.

BOUSQUET-VERNHETTES C. (2002), Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine – Décodage conceptuel stochastique, Thèse de doctorat de *l'université de Toulouse III*.

HADDAD B., YASEEN M. (2005), A Compositional Approach Towards Semantic Representation and Construction of ARABIC, Actes de *LACL*.

JAMOUSSI S. (2004), Méthodes statistiques pour la compréhension automatique de la parole, Thèse de doctorat de *l'université Henri Poincaré*.

Kohonen T. (1998), Self-organisation and associative memory. Berlin, Spring-Verlag.

McQueen J. (1967), Some methods for classification and analysis of multivariate observations, Actes de *the Berkeley Symposium on Mathematical Statistics and Probability*.

MEFTOUH K., LASKRI M.T. (2001), Generation of the Sense of a Sentence in Arabic Language with a Connectionist Approach, Actes de *AICCSA*.

MINKER W. (1999), *Compréhension automatique de la parole spontanée*, Paris, L'Harmattan.

OUERSIGHNI R. (2001), A major offshoot of the Dinar-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic texts, Actes de *ACL/EACL*.

Rabiner L.R., Juang B.H. (1986), Introduction to Hidden Markov Models, *IEEE Transactions on Acoustics, Speech and Signal processing*.

ROSENFELD R. (1994), Adaptive statistical language modelling: A maximum entropy approach., Thèse de doctorat de *l'université de Carnegie Mellon*.

Schwartz R., Miller S., Stallard D., Makhoul J. (1996), Language Understanding Using Hidden Understanding Models, Actes de *ICSLP*.

SENEFF S. (1992), Robust parsing for spoken language systems, Actes de *ICASSP*, 189-192.

Van Noord G., Bouma G., Koeling R., Nederhof M.J. (1999), Robust grammatical analysis for spoken dialogue systems, *Natural Language Engineering 5(1)*.

Villaneau J., Antoine J.Y., Ridoux O. (2001), Combining Syntax and Pragmatic Knowledge for the Understanding of Spontaneous Spoken Sentences, Actes de *LACL'01*.

Session
Démonstrations

Présentation du logiciel Antidote RX

Éric BRUNELLE, Simon CHAREST
Druide informatique inc.
1435, rue St-Alexandre, bureau 1040
Montréal (Québec) H3A 2G4, Canada
developpement@druide.com

Antidote RX est la sixième édition d'Antidote, un logiciel d'aide à la rédaction développé et commercialisé par la société Druide informatique. Antidote RX comporte un correcteur grammatical avancé, dix dictionnaires de consultation et dix guides linguistiques. Il fonctionne sous les systèmes d'exploitation Windows, Mac OS X et Linux.

1 Correcteur

Fondé sur un analyseur en dépendances, le correcteur d'Antidote transforme les résultats de son analyse en diagnostics utilisables par le grand public. Il propose des corrections d'orthographe, de syntaxe, d'accords grammaticaux, de conformité à l'usage, et plusieurs autres. Antidote RX corrige aussi certaines erreurs de nature sémantique, comme **perpétrer la tradition*, qui ne peuvent être repérées par la seule analyse syntaxique.

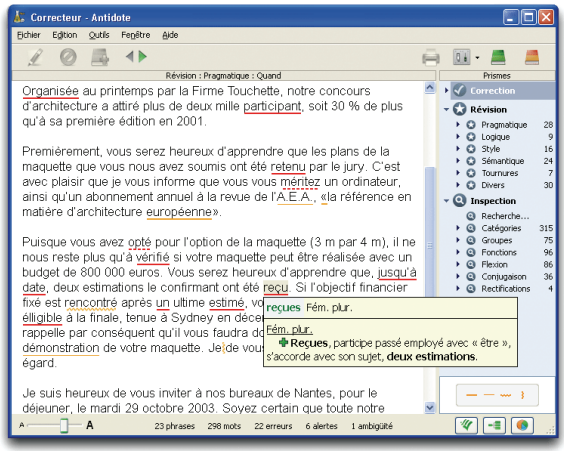


Figure 1 : le correcteur d'Antidote RX offre plus de 1 000 cas de détection et 25 000 corrections sémantiques.

2 Prismes de révision et d'inspection

En s'appuyant sur les résultats de l'analyse linguistique, le prisme de révision offre 32 filtres qui illustrent visuellement certains aspects pragmatiques (Qui, Quand, Où, Combien), logiques (charnières, citations), stylistiques (répétitions, verbes ternes, tournures passives et

négatives) et sémantiques (positif, négatif, fort, faible) d'un texte. Le prisme d'inspection, de son côté, en révèle les éléments constitutifs : catégories, groupes, fonctions et autres. En soumettant son texte à la correction puis aux filtres des prismes, l'utilisateur dispose d'une panoplie inédite d'outils pour l'examiner, le corriger, le réviser et le raffiner en profondeur.

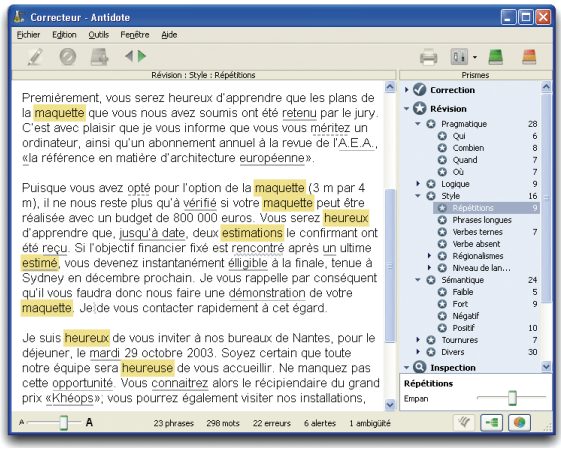


Figure 2 : le filtre des répétitions surligne les redites potentielles.

3 Dictionnaires

Antidote RX propose dix grands dictionnaires du français en une interface unifiée de consultation et de présentation. Des définitions aux synonymes, des cooccurrences aux analogies, les dictionnaires d'Antidote offrent une référence lexicale d'une richesse, d'une variété et d'une cohésion fort intéressantes.

Le dictionnaire de cooccurrences est une nouveauté de l'édition RX. Il a été constitué essentiellement automatiquement à partir d'un corpus de 500 millions de mots. L'analyseur syntaxique d'Antidote a été mis à contribution pour recenser plus de 17 millions de paires de mots liées par diverses relations syntaxiques. Les cooccurrences les plus significatives ont été dégagées à l'aide d'un filtre statistique et d'une révision manuelle. Le résultat est un dictionnaire de 800 000 cooccurrences illustrées par plus de 2 millions de phrases exemples tirées du corpus. À notre connaissance, il s'agit du plus vaste dictionnaire de cooccurrences du français à ce jour.

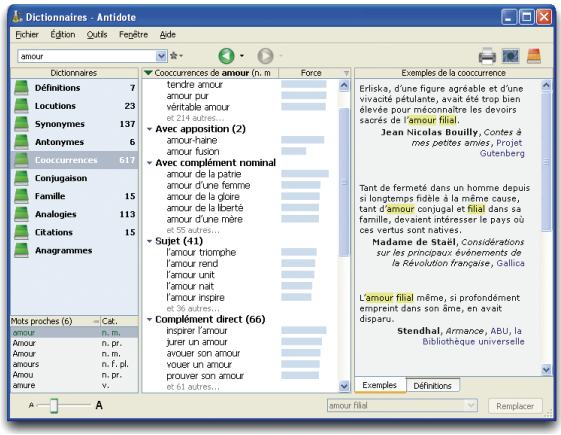


Figure 3 : le dictionnaire de cooccurrences compte 800 000 cooccurrences illustrées par plus de 2 millions de phrases exemples.

4 Guides linguistiques

L'utilisateur humain se contente rarement des diagnostics de la machine. Pour éclairer ses raisonnements, Antidote présente dix guides linguistiques totalisant 600 articles et couvrant tous les sujets pertinents à l'écriture du français. Lorsque le correcteur signale une erreur, il ouvre directement l'article correspondant pour que l'utilisateur puisse juger de la pertinence de sa détection.

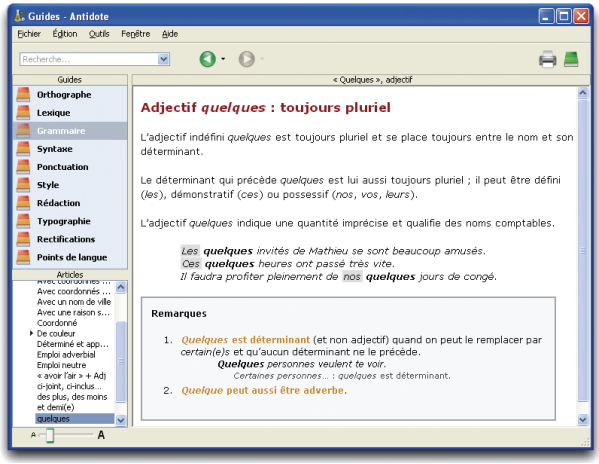


Figure 4 : le guide de grammaire explique les règles de l'orthographe grammaticale.

Logiciel Cordial

Dominique LAURENT, Sophie NÈGRE, Patrick SÉGUÉLA
Synapse Développement
33 Rue Maynard,
31000 Toulouse
http://www.synapse-fr.com

Résumé. Cordial est un correcteur efficace et discret enrichi d'un grand nombre de fonctions d'aide à la rédaction et d'analyse de documents. Très riche avec ces multiples dictionnaires et souvent pertinent dans ses propositions, Cordial est un compagnon précieux qui vous permet d'assurer la qualité de vos écrits. La version 2007 de Cordial s'intègre dans un vaste éventail de logiciels comme les traitements de texte (Word, Open Office, Word Perfect...), clients de messagerie (Outlook, Notes, Thunderbird, webmails...) ou navigateurs (Explorer, Mozilla).

1 Correcteur orthographique et grammatical multilingue

Cordial 2007 met en œuvre des milliers de règles de grammaire qui s'appuient sur des dictionnaires orthographiques (205 000 lemmes) et grammaticaux (104 000 lemmes) très pointus. Cordial s'appuie sur le sens des mots avant la correction. Il corrige ainsi près de 9 fautes sur 10 et évite de corriger des phrases justes ! Il atteint des scores exceptionnels : moins d'un faux message toutes les deux pages !

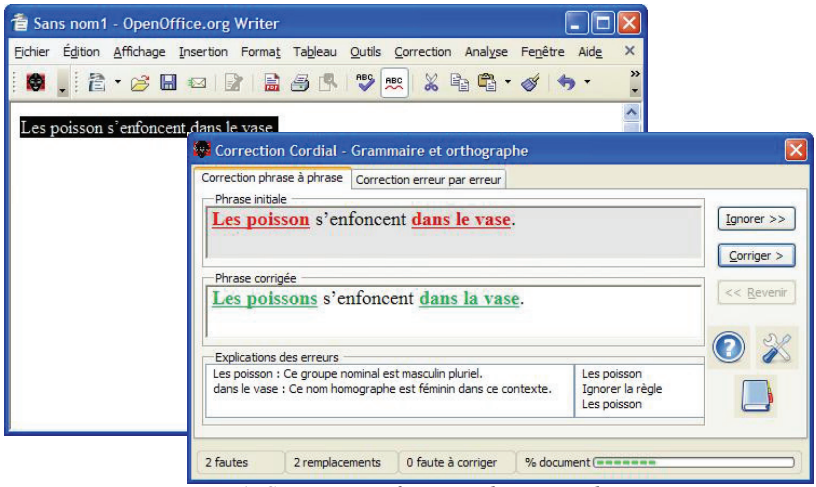


Figure 1. Correction en français, phrase par phrase.

2 Aides à la rédaction

Cordial est aussi un atelier d'aides à la rédaction intégrant les dictionnaires de références du français, noms propres et noms communs, Littré, Trésor de la Langue Française, dictionnaire de synonymie, dictionnaire des analogies et homonymies, traducteur mot-à-mot , conjugueur, aide grammaticale intelligente.

Ces ressources sont accessibles dans tous les logiciels ou sur simple clic via un mécanisme de Pop up.

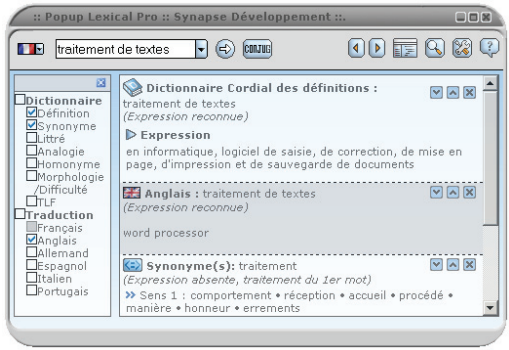


Figure 2. Pop up Lexical : une interface hypertextuelle pour naviguer dans les ressources dictionnairiques

Le dictionnaire de synonymie propose 4,5 millions de liens pondérés entre les sens de chaque mot mais également entre les expressions nominales et verbales.

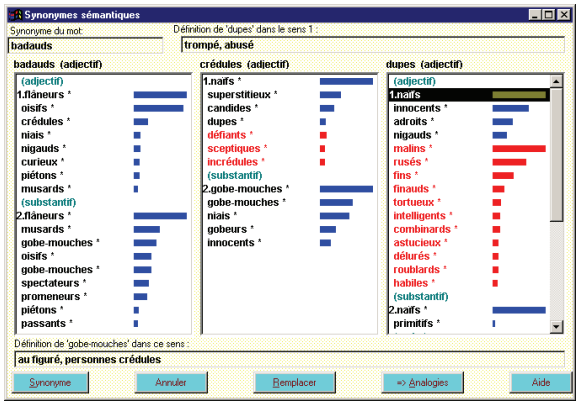
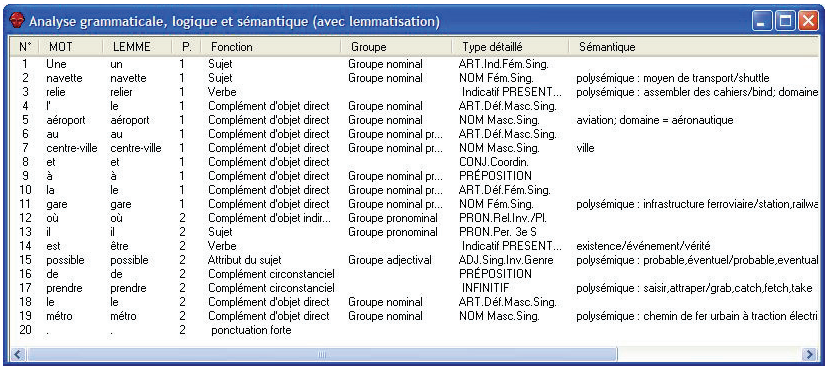


Figure 3. Dictionnaire des synonymes.

Cordial 2007 offre des traductions (mot à mot ou expression à expression) entre : français, anglais, allemand, espagnol, italien, portugais. Vous pouvez dans n'importe quel logiciel demander la traduction d'un mot ou d'une expression d'une de ces cinq langues vers une autre. Plus de 200 000 mots et expressions du français vers l'anglais et de l'anglais vers le français, au total plus de 2.5 millions de traduction de mots et expressions.

3 Analyseur syntaxique et sémantique

Enfin, Cordial propose des fonctions nombreuses et performantes qui font habituellement l'objet de logiciels spécialisés. Ces fonctions d'analyse de données textuelles et de synthèse de texte ont été développées et généralisées avec un soin particulier et une amplitude sans précédent dans les correcteurs grammaticaux habituels.



N°	MOT	LEMME	P.	Fonction	Groupe	Type détaillé	Sémantique
1	Une	un	1	Sujet	Groupe nominal	ART.Ind.Fém.Sing.	
2	navette	navette	1	Sujet	Groupe nominal	NOM Fém.Sing.	polysémique : moyen de transport/shuttle
3	relie	relier	1	Verbe		Indicatif PRESENT...	polysémique : assembler des cahiers/bind; domaine
4	l'	le	1	Complément d'objet direct	Groupe nominal	ART.Déf.Masc.Sing.	
5	aéroport	aéroport	1	Complément d'objet direct	Groupe nominal	NOM Masc.Sing.	aviation; domaine = aéronautique
6	au	au	1	Complément d'objet direct	Groupe nominal pr...	ART.Déf.Masc.Sing.	
7	centre-ville	centre-ville	1	Complément d'objet direct	Groupe nominal pr...	NOM Masc.Sing.	ville
8	et	et	1	Complément d'objet direct		CONJ.Coordin.	
9	à	à	1	Complément d'objet direct	Groupe nominal pr...	PRÉPOSITION	
10	la	le	1	Complément d'objet direct	Groupe nominal pr...	ART.Déf.Fém.Sing.	
11	gare	gare	1	Complément d'objet direct	Groupe nominal pr...	NOM Fém.Sing.	polysémique : infrastructure ferroviaire/station,railwa
12	où	où	2	Complément d'objet indir...	Groupe pronominal	PRON.Rel.Inv./Pl.	
13	il	il	2	Sujet	Groupe pronominal	PRON.Per. 3e S	
14	est	être	2	Verbe		Indicatif PRESENT...	existence/événement/vérité
15	possible	possible	2	Attribut du sujet	Groupe adjectival	ADJ.Sing.Inv.Genre	polysémique : probable,éventuel/probable,eventual
16	de	de	2	Complément circonstanciel		PRÉPOSITION	
17	prendre	prendre	2	Complément circonstanciel		INFINITIF	polysémique : saisir,attraper/grab,catch,fetch,take
18	le	le	2	Complément d'objet direct	Groupe nominal	ART.Déf.Masc.Sing.	
19	métro	métro	2	Complément d'objet direct	Groupe nominal	NOM Masc.Sing.	polysémique : chemin de fer urbain à traction électri
20	.	.	2	ponctuation forte			

Figure 4. Analyse morphologique, syntaxique et sémantique.

L'analyse syntaxique de Cordial, pierre angulaire des développements chez Synapse Développement, a été récemment désignée comme l'une des plus pertinentes pour le français, lors de la campagne Easy du programme Technolangue, où elle a obtenu la F-Measure globale la plus élevée. Cette analyse est également disponible dans le logiciel *Cordial Analyseur*.

Dans Cordial, l'analyse sémantique est utilisée dans la plupart des traitements. Une fonction reprend dans une représentation « en éclaté » les informations thématiques et les concepts clés (en blanc) trouvés sur un texte. D'un seul coup d'œil, vous pouvez ainsi voir quels sont les thèmes présents dans le texte. Voici comment elle se présente sur ce texte.

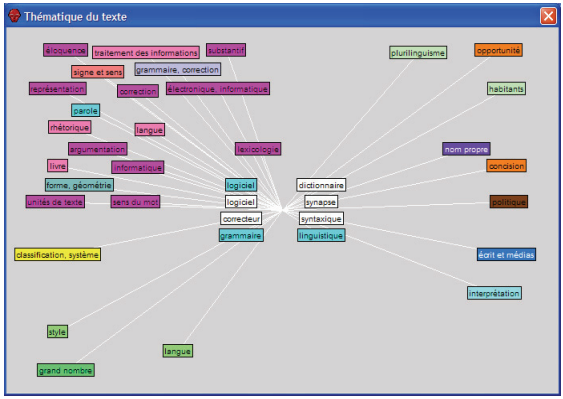


Figure 5. Analyse thématique de ce texte.

Toutes les briques technologiques disponibles dans Cordial sont disponibles sous la forme de composants indépendants, sous Windows et Linux. Ceci permet l'intégration directe de ces technologies de traitement du langage dans les chaînes de traitement.

TransCheck : un vérificateur automatique de traductions

Elliott MACKLOVITCH, Guy LAPALME
Laboratoire RALI – Université de Montréal
Montréal, Canada
{macklovi,lapalme}@iro.umontreal.ca

Résumé. Nous offrirons une démonstration de la dernière version de TransCheck, un vérificateur automatique de traductions que le RALI est en train de développer. TransCheck prend en entrée deux textes, un texte source dans une langue et sa traduction dans une autre, les aligne au niveau de la phrase et ensuite vérifie les régions alignées pour s'assurer de la présence de certains équivalents obligatoires (p. ex. la terminologie normalisée) et de l'absence de certaines interdictions de traduction (p. ex. des interférences de la langue source). Ainsi, TransCheck se veut un nouveau type d'outil d'aide à la traduction qui pourra à réduire le fardeau de la révision et diminuer le coût du contrôle de la qualité.

Abstract. We will present a demonstration of the latest version of *TransCheck*, an automatic translation checker that the RALI is currently developing. *TransCheck* takes as input two texts, a source text in one language and its translation in another, aligns them at the sentence level and then verifies the aligned regions to ensure that they contain certain obligatory equivalents (e.g. standardized terminology) and do not contain certain prohibited translations (e.g. source language interference). *TransCheck* is thus intended to be a new type of tool for assisting translators which has the potential to ease the burden of revision and diminish the costs of quality control.

Mots-clés : traduction assistée par ordinateur, vérification automatique de traductions, révision de traduction.

Keywords: machine-aided translation, automatic translation checking, translation revision.

En collaboration avec le Centre de recherche en technologies langagières à Ottawa, le laboratoire RALI est en train de développer un nouveau prototype de *TransCheck*, un vérificateur automatique de traductions. Comme son nom l'indique, *TransCheck* est un outil qui vise à aider un réviseur humain à déceler automatiquement certaines erreurs de traduction et à faire respecter certaines normes linguistiques au sein d'un grand service de traduction. Le système ressemble un peu à un vérificateur d'orthographe, dans le sens qu'il sert à repérer des erreurs potentielles dans un texte. Mais contrairement à un vérificateur d'orthographe, qui décèle des erreurs de forme monolingues, *TransCheck* cible des erreurs de correspondances entre deux textes, un texte source et sa traduction. Le mot « librairie », par exemple, est une forme correcte en français, mais il constitue néanmoins une erreur en tant que traduction du mot anglais « library ». Un vérificateur monolingue serait muet devant de tels cas, alors qu'un vérificateur bi-textuel comme *TransCheck* sera en mesure de le signaler.

Conceptuellement, comment fonctionne ce vérificateur automatique de traductions ? *TransCheck* prend en entrée deux textes, un texte source dans une langue et une ébauche de traduction dans une autre. Dans un premier temps, le système segmente chaque texte en unités linguistiques (c.-à-d. des phrases et des mots), pour ensuite les aligner au niveau des phrases. Une fois qu'il a établi cette correspondance entre les phrases qui sont des traductions mutuelles, *TransCheck* y applique les différents modules de détection d'erreurs. Certains de ces modules sont basés sur des grammaires internes, comme les grammaires des différents types d'expressions numériques (p. ex. les dates ou les expressions monétaires) ; d'autres modules exigent des données externes fournies normalement par l'utilisateur (p. ex. la terminologie). *TransCheck* signale une erreur potentielle lorsque, dans une région alignée, il repère un item source sans trouver son correspondant obligatoire, ou lorsque le système repère un item source et trouve aussi son correspondant interdit dans la région cible.

Nous pouvons illustrer ces différents cas de figures à l'aide de la capture d'écran à la page suivante. Comme on peut y voir, le prototype actuel est pourvu d'une interface graphique dans laquelle le texte source apparaît à gauche et le texte cible à droite. Les cases à cocher au dessus des deux textes permettent à l'utilisateur d'activer ou de désactiver les différents modules de détection d'erreurs. Les régions alignées sont démarquées par une ligne horizontale bleue et, au sein de ces régions, les erreurs décelées sont signalées par un soulignement rouge. Dans la troisième région, par exemple, où l'utilisateur a pointé sur le mot souligné « sniper », le système signale une erreur potentielle de terminologie ; et, dans le cadre inférieur de la fenêtre, il indique les termes français qu'il cherchait mais n'a pas trouvés du côté cible. Dans la région suivante, il s'agit d'un autre type de correspondance obligatoire : une expression numérique cette fois, qui se trouve à être traduit par la paraphrase « un tour complet ». Dans la dernière région alignée, nous voyons un exemple d'une interdiction de traduction : normalement, le nom « deception » en anglais et son homographe français sont considérés comme étant des faux-amis.

Il ne faut pas s'imaginer qu'un outil comme *TransCheck* pourra déceler toutes les erreurs de traduction, ce qui représente un problème qui est probablement plus difficile que la traduction automatique elle-même. Par contre, un système qui serait en mesure d'alléger le fardeau des réviseurs en automatisant les aspects les plus mécaniques et fastidieux de leur travail serait certainement bien accueilli. Ainsi, le prototype actuel traite un sous-ensemble d'erreurs que l'on peut détecter par des techniques simples et formelles, c.-à-d. sans 'compréhension profonde' des textes. Ceci dit, *TransCheck* n'est pas du tout un système fermé ; au fur et à mesure que notre compréhension des relations traductionnelles avancera, nous pourrons y intégrer d'autres types de vérifications.

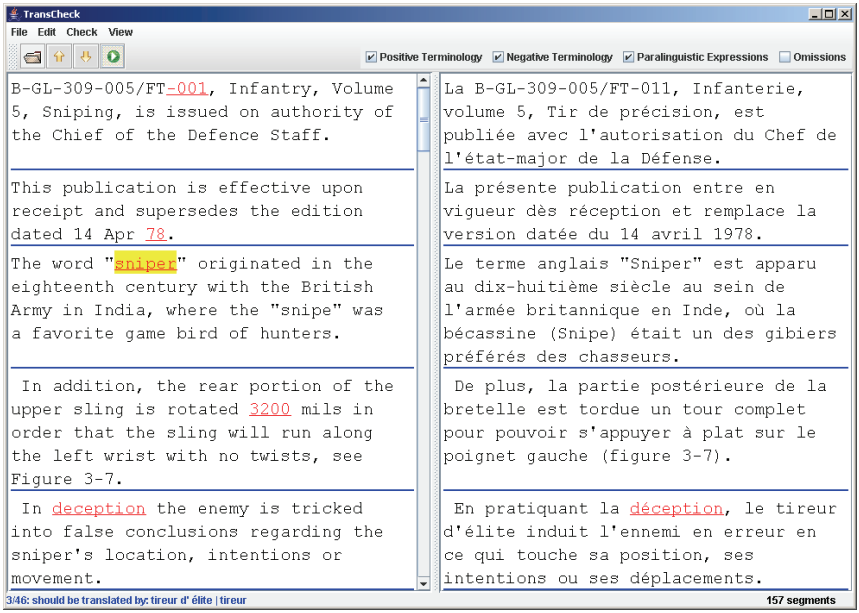


Figure 1 : Une capture d'écran du système *TransCheck*
où les erreurs signalées par le système sont soulignées en rouge.

Finalement, il est important de noter que *TransCheck* est un outil d'aide à la traduction qui vise surtout à appuyer le traducteur et non pas à le remplacer. Ainsi, le système signale des erreurs potentielles, mais c'est à l'utilisateur de décider lesquelles doivent réellement être corrigées et lesquelles peuvent rester telles quelles. Autre élément interactif dans l'emploi de *TransCheck*, l'utilisateur peut aisément modifier ou ajouter des entrées au glossaire de la terminologie et à l'anti-dictionnaire des interdictions dont *TransCheck* se sert pour effectuer ces vérification.

Références

JUTRAS J-M. (2000). An Automatic Reviser : The TransCheck System. *Sixth Applied Natural Language Processing Conference,* 127-134.

MACKLOVITCH E. (1995). TransCheck - or the Automatic Validation of Human Translations. *Proceedings of MT Summit V (Luxembourg).*

RUSSELL, G. (2005). Automatic Detection of Translation Errors: the TransCheck System. *Translating and the Computer 27 (ASLIB Conference).*

Le CNRTL, Centre National de Ressources Textuelles et Lexicales, un outil de mutualisation de ressources linguistiques

Jean-Marie PIERREL, Etienne PETITJEAN
CNRTL/ATILF CNRS – Nancy Université
44 avenue de la Libération, BP 30687, 54063 Nancy CEDEX
{Jean-Marie.Pierrel,Etienne.Petitjean}@atilf.fr

Résumé. Créé en 2005 à l'initiative du Centre National de la Recherche Scientifique, le CNRTL propose une plate-forme unifiée pour l'accès aux ressources et documents électroniques destinés à l'étude et l'analyse de la langue française. Les services du CNRTL comprennent le recensement, la documentation (métadonnées), la normalisation, l'archivage, l'enrichissement et la diffusion des ressources. La pérennité du service et des données est garantie par le soutien institutionnel du CNRS, l'adossement à un laboratoire de recherche en linguistique et informatique du CNRS et de Nancy Université (ATILF – Analyse et Traitement Informatique de la Langue Française), ainsi que l'intégration dans le réseau européen CLARIN (common language resources and technology infrastructure european).

1 Les missions du CNRTL

Les missions du CNRTL (www.cnrtl.fr) mis en place par le CNRS, Département Sciences de l'Homme et de la Société et Direction de l'Information Scientifique, au sein de l'ATILF peuvent se résumer en sept points :

- « Entrées » : acceptation, contrôle et validation des ressources, tant d'un point de vue scientifique que technique, afin d'assurer la qualité des ressources (corpus dictionnaires, lexiques et outils de traitement) offertes par le centre ;
- « Stockage » : stockage, maintenance et récupération des ressources. Beaucoup de chercheurs et d'équipes en SHS qui développent pour leurs recherches propres des ressources informatisées ne disposent en effet pas des moyens nécessaires pour assurer cette fonction ;
- « Gestion des ressources » : partage, conservation et enrichissement de ressources, afin d'assurer une réelle mutualisation entre équipes de recherche ;
- « Administration » : administration des ressources et aide aux utilisateurs ;
- « Pérennisation et documentation » : mise à jour et évolution des supports informatiques. L'évolution des matériels et logiciels informatiques nécessite une maintenance régulière de telles ressources informatisées pour éviter des gâchis que nous avons pu connaître dans le passé où certains corpus ont été perdus par manque de maintenance et de pérennisation ;
- « Accès » : aide et réponse aux utilisateurs permettant aux non spécialistes de l'informatique que sont les chercheurs en SHS d'accéder et d'exploiter au mieux de telles ressources informatisées à travers des outils adaptés à leurs besoins ;
- « Formation » : formation des producteurs et utilisateurs aux méthodologies d'annotation, de codage et de normalisation. Sur ce point fort du fait que Nancy est centre support de la TEI, on s'appuie autant que faire se peut sur les recommandations de la TEI.

2 Le CNRTL nœud d'un réseau international européen

Au-delà de sa seule mission nationale, le CNRTL participe au réseau européen CLARIN (Common Language Resource and Technologie Infrastructure : http://www.mpi.nl/clarin) des centres de gestion de ressources linguistiques qui correspond à l'une des propositions européennes d'infrastructure de recherche en SHS incluse dans la feuille de route ESFRI qui définit les infrastructures de recherches à soutenir dans le cadre du 7ème programme-cadre. Elle vise à définir une infrastructure européenne partagée par les grands centres de recherche européens et s'appuyant sur des centres régionaux « certifiés » dans leurs domaines respectifs.

Ce projet est également l'occasion d'organiser une réflexion commune sur la gestion d'une plate-forme ouverte de gestion et d'archivage de documents numériques avec nos collègues du Max Planck Institute qui travaillent actuellement sur le même sujet. Dans l'idéal, cette collaboration permettra de converger vers une plate-forme logicielle unique utilisable par le CNRTL comme par le MPI. Cette plate-forme logicielle pourrait s'articuler autour de Fedora (http://www.fedora.info/) qui est un projet open-source offrant une architecture flexible pour la gestion et la distribution de documents numériques. Développé conjointement par l'université de Virginie et l'université de Cornell, ce système semble offrir les bases dont nous avons besoin pour développer cette plate-forme, à savoir :

- Le dépôt de ressources : permettre à un utilisateur de pouvoir soumettre une ou plusieurs ressources numériques (texte brut ou étiqueté morpho-syntaxiquement, etc.)
- La consultation des ressources : offrir aux utilisateurs une interface de consultation permettant la navigation et la sélection des différents corpus et ressources disponibles sur la plateforme.
- Le téléchargement des ressources : faciliter le téléchargement des ressources sélectionnées dans le format de sortie souhaité par les utilisateurs (XML, PDF, Word, HTML, etc.)

3 Les ressources accessibles au sein du CNRTL

Le CNRTL s'est structuré autour de cinq pôles de compétence : un portail lexical sur le français ; des corpus et données textuelles, annotés ou non ; des dictionnaires encyclopédiques et linguistiques (anciens et modernes) ; des lexiques phonétiques, morphologiques, syntaxiques, sémantiques ; des outils linguistiques (étiqueteurs, analyseurs, aligneurs, concordances, outils d'annotation).Afin de proposer une première offre de ressources au sein du CNRTL, nous avons travaillé dans un premier temps sur la base des ressources linguistiques informatisées actuellement disponibles à Nancy, ressources qui, suivant les cas, sont des ressources libres et téléchargeables après acceptation d'une licence de type ressources libres, des ressources sous droits accessibles uniquement via une interface web spécifique, ressources sous droits accessibles uniquement dans le cadre d'une convention de partenariat avec les ayants droits. Parmi les ressources déjà intégrées au CNRTL, outre les outils et le portail lexical sur lesquels nous allons reveni, il convient de noter :

Les corpus de textes libres de droit d'auteur et d'éditeur (dans un premier temps 500 textes issus de FRANTEXT) : à travers une sélection par auteurs, titres, dates ou genres, nous offrons la possibilité de télécharger les textes sélectionnés au format XML dans une DTD respectant les recommandations de la TEI : l'utilisateur récupère une archive contenant la DTD et le codage XML/TEI des textes (à notre connaissance, le CNRTL est le premier site offrant un ensemble de corpus français téléchargeables et normalisés XML/TEI d'environ 150 millions de caractères) ; et un corpus annoté pour le traitement des DEscriptions DEfinies (DEDE : coopération LORIA, Metadif et ATILF).

Le lexique Morphalou en accès libre tant en consultation qu'en téléchargement : lexique ouvert des formes fléchies du français qui fournit 524 725 formes fléchies, appartenant à 95 810 lemmes, linguistiquement valides (responsabilité d'un comité éditorial) et respectant les propositions de normalisation pour les ressources lexicales de l'ISO (TC37/SC4).
Des dictionnaires tant modernes qu'anciens : outre l'accès à la version électronique du Trésor de la Langue Française, dictionnaire de référence des 19e et 20e siècles, produit par le laboratoire ATILF, un ensemble de liens permet également la consultation des dictionnaires suivants : le Dictionarium latinogallicum (troisième édition - 1552) de Robert Estienne, le Thresor de la langue françoyse, tant ancienne que moderne de Jean Nicot (Paris, David Douceur, 1606), le Dictionnaire historique et critique de Bayle (fac-similé de la version de 1740), le Dictionnaire critique de la langue française de Jean-François Féraud (1787-1788), le Dictionnaire de l'Académie française (1e édition 1694 - 4e édition 1762 - 5e édition 1798 - 6e édition 1835 - 8e édition 1932/1935 – 9e édition en cours)

4 Des outils à disposition de la communauté

Le CNRTL se propose également de mettre à disposition de la communauté des outils linguistiques utilisables directement sur le site Web à partir d'un simple navigateur Internet. Parmi les différents projets en cours ou à venir, nous comptons offrir aux utilisateurs un accès simple et convivial à des outils comme :

- FLEMM : outil d'analyse flexionnelle de textes en français qui ont été au préalable été étiqueté, au moyen de l'un des deux catégorisateurs : Brill ou TreeTagger.
- DERIF : outil d'analyse morpho-sémantique du français qui s'applique à des entrées lexicales catégorisées issues d'un dictionnaire de la langue générale, capable de traiter des mots hors-dictionnaire et dont les résultats associent la morphologie et la sémantique
- POMPAMO : outil de détection de candidats à la néologie formelle et catégorielle basé sur l'utilisation de lexiques d'exclusion. Ce projet exploite des ressources lexicales comme Morphalou et permet d'en constituer de nouvelles.

5 Un exemple d'intégration de ressources : le portail lexical

Le portail lexical, quant à lui, a pour vocation de valoriser et de partager un ensemble de données issues des travaux de recherche sur le lexique français. Projet évolutif, cette base de connaissances lexicales exploite aujourd'hui divers documents numériques pour fournir, à partir d'une forme lexicale, six types d'informations importantes : des informations morphologiques issues de Morphalou (www.atilf.fr/morphalou), des informations lexicographiques et étymologiques issues des projets TLF (www.atilf.fr/tlfi) et TLF-Etym, des informations de synonymies à travers l'intégration du dictionnaire de synonymes de Caen (http://www.crisco.unicaen.fr/), une concordance utilisant le corpus des textes de la base Frantext (www.atilf.fr/frantext) et une présentation des résultats de proxémie du projet Prox de l'ERSS (http://w3.univ-tlse2.fr/erss/). Il offre aussi la possibilité d'exporter les résultats du concordancier au format XML/TEI. C'est à notre connaissance le seul site permettant à un utilisateur d'exporter dans un format normalisé un concordancier français d'une telle importance. Ces informations sont directement intégrables dans d'autres applications Web à travers des liens spécifiques à chacune des formes type d'informations tels que : www.cnrtl.fr/concordance/*mot*. De plus, le portail lexical permet une hyper-navigation vers toutes les informations lexicales disponibles pour ce mot. Par exemple, si l'on veut obtenir des informations sur un mot d'un exemple de concordance, un double-clic sur le mot affiche un menu qui permet d'hyper-naviguer vers les informations lexicales de ce mot.

RÉCITAL-2007

5 au 8 juin 2007, Toulouse, France

Actes de la 11e RENCONTRE

DES ÉTUDIANTS CHERCHEURS EN INFORMATIQUE

POUR LE TRAITEMENT AUTOMATIQUE DES LANGUES

(communications affichées)

Éditeurs scientifiques

Farah BENAMARA et Sylwia OZDOWSKA

Organisation de la conférence

CLLE-ERSS (UMR 5263) & IRIT (UMR 5505)

Sous l'égide de l'ATALA

(Association pour le Traitement Automatique des Langues)

Comité d'organisation

Nathalie AUSSENAC-GILLES	*(CNRS, IRIT)*
Farah BENAMARA*	*(Université Paul Sabatier, IRIT)*
Jean-Léon BOURAOUI	*(Université Paul Sabatier, IRIT)*
Didier BOURIGAULT	*(CNRS & Université Toulouse Le Mirail, CLLE)*
Véronique DEBATS	*(CNRS, IRIT)*
Fabrice ÉVRARD	*(Institut National Polytechnique, IRIT)*
Cécile FABRE	*(Université Toulouse-le-Mirail, CLLE)*
Edith GALY	*(Université Toulouse-le-Mirail, CLLE)*
Bruno GAUME	*(Université Toulouse-le-Mirail, CLLE)*
Nabil HATHOUT	*(CNRS, Université Toulouse Le Mirail, CLLE)*
Dominique LONGIN	*(CNRS, IRIT)*
Josiane MOTHE	*(Université Paul Sabatier, IRIT)*
Philippe MULLER	*(Université Paul Sabatier, IRIT)*
Sylwia OZDOWSKA*	*(Université Toulouse-le-Mirail, CLLE)*
Patrick SAINT-DIZIER	*(CNRS, IRIT)*
Frank SAJOUS	*(Université Toulouse-le-Mirail, CLLE)*
Ludovic TANGUY	*(Université Toulouse-le-Mirail, CLLE)*
Laure VIEU	*(CNRS, IRIT)*

Comité de programme

Jean-Yves ANTOINE	*(Université de Tours, LI)*
Frédéric BECHET	*(Université Avignon, LIA)*
Farah BENAMARA*	*(Université Paul sabatier, IRIT)*
Laurent BESACIER	*(Université Joseph Fourrier, CLIPS IMAG)*
Hervé BLANCHON	*(Université Pierre Mendès-France, CLIPS IMAG)*
Philippe BOULA DE MAREUIL	*(CNRS, LIMSI)*
Estelle CAMPIONE	*(Université de Provence, DELIC)*
Vincent CLAVEAU	*(Université Rennes 1, IRISA)*
Cécile FABRE	*(Université Toulouse-Le-Mirail, CLLE)*
Thierry HAMON	*(Université Paris 13, LIPN)*
Philippe LANGLAIS	*(Université de Montréal, RALI)*
Fabrice MAUREL	*(Université de Caen, LMNO)*
Emmanuel MORIN	*(Université de Nantes, LINA)*
Alexis NASR	*(Université Paris 7, LATTICE)*
Sylwia OZDOWSKA*	*(Université Toulouse-le-Mirail, CLLE)*
Thierry POIBEAU	*(Université Paris 13, LIPN)*
Laurent ROUSSARIE	*(Université Paris 8, LATTICE)*
Ludovic TANGUY	*(Université Toulouse-Le-Mirail, CLLE)*

* Présidente

Session
Communication affichées

Vers une nouvelle structuration de l'information extraite automatiquement

Alejandro ACOSTA
LATTICE-CNRS (UMR 8094)
Université Paris 7
alejandro.acosta@linguist.jussieu.fr

Résumé. Les systèmes d'Extraction d'Information se contentent, le plus souvent, d'enrichir des bases de données plates avec les informations qu'ils extraient. Nous décrivons dans cet article un travail en cours sur l'utilisation de données extraites automatiquement pour la construction d'une structure de représentation plus complexe. Cette structure modélise un réseau social composé de relations entre les entités d'un corpus de biographies.

Abstract. Information Extraction systems are widely used to create flat databases of templates filled with the data they extract from text. In this article we describe an ongoing research project that focuses on the use of automatically extracted data to create a more complex representation structure. This structure is a model of the social network underlying the relations that can be established between the entities of a corpus of biographies.

Mots-clés : extraction d'information, analyse de réseaux sociaux, biographies, entités nommées, représentation de connaissances.

Keywords: information extraction, social network analysis, named entities, biographies, knowledge representation.

1 Introduction

Les systèmes de Traitement Automatique de Langues (TAL) sont nés, en grande partie, du désir de modéliser la compréhension du langage naturel et de créer des systèmes capables de transformer un discours incompréhensible pour une machine en une représentation formelle explicite de l'information véhiculée par le langage. Cette tâche s'est avérée plus difficile que prévu et en conséquence les traducteurs automatiques, ainsi que les systèmes intelligents autonomes restent du domaine de la science-fiction.

Le domaine de la Représentation de Connaissances s'est concentré sur le problème de la structuration des contenus, et sur les opérations sur les structures de représentation. De son côté, la Compréhension du Discours s'est heurtée à la complexité de l'interprétation du langage naturel et a dû chercher à être plus modeste dans ses objectifs.

L'Extraction d'Information (EI) est née du désir de construire des systèmes capables de répondre à des tâches spécifiques de compréhension. Son évolution est fortement liée aux campagnes d'évaluation de systèmes capables de repérer des événements ponctuels, ainsi que les

acteurs associés à ces événements. Les évaluations ont encouragé les avancées dans le domaine, mais les techniques d'évaluation ont aussi influencé la définition d'un domaine qui, à l'origine, était celui de la compréhension des textes. Cette influence a dirigé l'EI vers des applications où l'on cherche à repérer des faits spécifiques dans les textes d'un domaine particulier. Ces systèmes sont capables de collecter un ensemble d'informations, structurées dans des formulaires qui remplissent des bases de données plates.

Ces bases de données peuvent être utilisées facilement pour l'évaluation des systèmes en termes de rappel et de précision (et de leurs mesures dérivées). Or, il est plus difficile d'évaluer des structures plus complexes, mettant en rapport les événements entre eux, ou les acteurs qui se répètent dans les différents événements (une représentation plus proche des bases de données relationnelles qui sont devenues le standard). De ce fait, on s'est contenté de construire des systèmes capables de faire ce qui était nécessaire pour leur évaluation et qui n'ont pas cherché à construire des structures de représentation plus riches.

Nous pensons que les techniques d'EI peuvent être utilisées pour construire des représentations plus complexes, des structures relationnelles. La volonté de structurer l'information dans des formulaire isolées a joué un rôle dans le choix des méthodes qui permettaient d'atteindre ce but. De la même manière, l'objectif de constituer une représentation plus complexe va aussi avoir une influence sur la définition de la tâche d'extraction.

Nous présentons dans cet article un travail en cours qui propose d'utiliser autrement les résultats d'une tâche d'extraction. Plus précisément, nous proposons une structure de représentation plus riche que celle qui est suggérée dans les tâches conventionnelles. Pour ce faire, nous présentons un travail mené sur un corpus qui se prête bien à la conception d'une représentation structurée : une vaste collection de notices biographiques.

Le reste de ce document est organisé de la manière suivante : dans la section 2 nous présentons le corpus de biographies, dans la section 3 nous discutons les motivations pour chercher une représentation plus complexe de l'information extraite, nous décrivons les éléments qui composent cette représentation (3.1 et 3.2) et nous parlons de l'état de l'art (3.3). Dans la section 4 nous présentons les techniques d'EI utilisées. Finalement, nous présentons quelques perspectives pour la continuation de ce travail dans la section 5.

2 Le Maitron

Le corpus utilisé dans ce travail est une collection de notices biographiques. Ces biographies peuvent être étudiées séparément mais il est aussi intéressant d'étudier l'objet plus large, et plus complexe, dont elles font partie.

Le *Dictionnaire biographique du mouvement ouvrier francais*[1] (dorénavant *Le Maitron*) est un dictionnaire contenant des notices biographiques sur les vies de milliers de personnes ayant participé activement aux luttes sociales de l'histoire française depuis 1789. Le dictionnaire original est divisé en quatre sections, correspondant à quatre périodes historiques qui vont de la Révolution Française au début de la Deuxième Guerre mondiale. La publication a commencé à s'enrichir d'une nouvelle section en 2006 avec l'ajout d'une nouvelle période entre la Deuxième Guerre et 1968.

[1] `http://www.maitron.org`

Plus de 600 auteurs ont participé au projet depuis sa création par Jean Maitron dans les années cinquante. Les notices biographiques des 56 volumes du dictionnaire (plus de 100 000 articles) sont constamment révisées et retravaillées par une équipe d'historiens dirigée par Claude Pennetier au Centre d'histoire sociale du XX^{e} siècle[2]. Le corpus du *Maitron* compte ainsi plus de 18 millions de mots, couvrant une période qui s'étend sur plus de deux siècles d'histoire.

Notre objectif est de proposer une représentation qui rende compte de la complexité du contenu du corpus (sans qu'elle soit exhaustive pour autant), et qui permette d'exploiter ces informations sous une autre forme.

3 Structuration des informations extraites

Les différentes approches traitant de la représentation de connaissances ont proposé un grand nombre de modèles dans l'étude des structures qui peuvent représenter symboliquement des informations complexes. Par ailleurs ces approches ont également proposé des techniques visant l'utilisation de ces structures pour le raisonnement automatique. Cependant, on est toujours incapable de traduire automatiquement le langage naturel en une représentation des connaissances qu'il transmet. On est malheureusement encore loin d'atteindre le niveau de performance que l'on espérait dans le domaine de la compréhension des textes il y a quelques décennies.

Le domaine de l'EI, de son côté, s'est contenté de la tâche qui consiste à récolter l'information dans des bases de données (Pazienza, 1999). Les *Message Understanding Conferences* de la fin des années 80 et des années 90 ont vu la transformation progressive de cette compréhension de messages en remplissage de descriptions d'événements (Poibeau, 2003).

l'EI est née pour affronter les difficultés rencontrées par les systèmes de compréhension des textes et chercher un compromis entre les limitations des systèmes et la qualité des résultats que l'on pouvait obtenir. La compréhension s'est vue réinterprétée et divisée en tâches successives, dont le but ultime était le remplissage de formulaires. Dans cette optique, chaque formulaire rempli correspond à une assertion sur le monde (on extrait l'information qu'on suppose vraie) ; l'ensemble de ces assertions constitue une base de données d'informations sur le domaine. Les systèmes d'EI se sont depuis penchés sur l'exploitation et le traitement d'importantes quantités de textes désormais disponibles en format numérique (et qui peuvent donc facilement être traités automatiquement).

Nous pensons que le compromis entre les limites des techniques et la complexité de la représentation peut encore être négocié. Certes, l'interprétation automatique du langage naturel, qui nous permettrait de constituer des bases de connaissances expressives et robustes, est encore hors de notre portée. Nous pensons cependant que les résultats d'un processus d'EI pourraient être améliorés en créant une sortie structurée de manière plus riche que la seule énumération des informations dans une base de données.

A notre avis, la structuration des informations pourrait être conçue autour des relations entre entités (dans le sens, grosso modo, d'entités nommées). Nous cherchons à établir des relations entre entités, plutôt qu'à repérer des événements pour remplir des formulaires avec des détails sur ces événements.

Le modèle de représentation que nous avons choisi pour traiter les informations extraites du

[2]CNRS / Université Paris I

Maitron s'inspire d'une technique de modélisation issue des sciences humaines : l'analyse des réseaux sociaux.

L'Analyse de Réseaux Sociaux (ARS) est une technique utilisée dans plusieurs domaines de recherche en sciences humaines pour étudier les relations entre individus. Ces individus sont souvent en rapport du fait de leur appartenance à un groupe ou une organisation. L'ARS s'est développée avec l'étude des rapports et interactions entre les individus de communautés particulières comme les habitants d'une île, les employés d'une entreprise ou les membres d'une bande d'adolescents d'un quartier, pour citer quelques exemples.

Avec l'avènement de l'ère de l'information et des communautés dites virtuelles de l'Internet, l'ARS attire l'attention d'une communauté scientifique de plus en plus large. Grâce à cette évolution du domaine, les chercheurs en sciences humaines, qui ont l'habitude de mener leurs recherches sur des corpus de taille limitée, commencent à profiter des exploits d'autres disciplines qui leur permettent d'étudier les relations entre individus à un autre niveau.

La théorie des graphes, pour citer un exemple, a développé des méthodes de représentation de réseaux et des techniques d'exploration qui sont maintenant utilisées pour interpréter les données relationnelles.

Le remplissage de formulaires avec des données biographiques n'est pas, en soi, une tâche qui permette d'enrichir l'information qui est déjà organisée dans notre corpus. En revanche, nous pouvons construire une structure à partir des nombreux liens entre les personnages du *Maitron*, une structure qui pourra être exploitée par les chercheurs qui utilisent le dictionnaire comme ressource de recherche.

Les sections 3.1 et 3.2 sont consacrées à la description des structures utilisées pour modéliser l'information contenue dans le corpus du *Maitron*. Nous considérons qu'une modélisation inspirée de l'ARS nous permet d'atteindre un niveau de structuration plus riche que celui inspiré du remplissage de formulaires d'extraction.

3.1 Le réseau social du *Maitron*

Si l'on voit le *Maitron* non pas comme une collection de notices biographiques mais comme un corps structuré d'informations à propos d'un ensemble d'individus, sa structuration sous la forme d'un réseau social apparaît comme une organisation presque naturelle pour ces données. En effet, le Maitron est composé d'histoires de vie d'individus qui étaient souvent en rapport les uns avec les autres, et c'est sur ces rapports que s'est bâtie une partie importante de l'histoire sociale de France. On remarquera que l'on peut associer ces individus à d'autres « acteurs de l'histoire » (publications, partis politiques, etc.). Les vies de ces personnes se rencontrent à l'intérieur d'événements spécifiques, à travers leur appartenance à des groupes ou à des associations. Les liens, une fois explicités, deviennent une abstraction structurée de l'interaction entre les objets du réseau.

Une structuration de ce type imite aussi une partie de la démarche de l'historien chercheur qui se sert de ce type de ressource documentaire : la première tâche du chercheur consistant à suivre les liens entre individus, associations, événements et autres entités saillantes. Individuellement, des données biographiques plus traditionnelles comme les dates de naissance, les lieux de naissances et autres, sont toujours des attributs porteurs de sens mais secondaires dans la structuration du corpus comme un tout.

On peut orienter le processus d'extraction vers la création d'une représentation qui aurait la structure d'un réseau. Le réseau peut être modélisé comme un graphe et la tâche d'extraction consiste à récolter les données nécessaires pour construire cette représentation (les liens entre les sommets du graphe). Tout comme le résultat d'une tâche classique d'EI le lien entre deux sommets est porteur de sens en tant qu'assertion d'un fait de l'univers (l'univers du corpus). En revanche, le sens du lien est plus riche car, dans le réseau, le même lien peut être interprété aussi en fonction de son appartenance à une structure plus complexe.

Sachant donc que nous cherchons à construire une représentation sous la forme d'un réseau, il nous reste à décrire les objets qui le composent : les sommets et les arcs de cette structure.

3.2 L'ontologie du *Maitron*

Nous avons décidé de structurer les objets qui composent notre représentation, le réseau, dans une ontologie. Très utilisées dans de nombreuses applications visant le passage du Web actuel au Web Sémantique, les ontologies permettent de modéliser les connaissances d'un domaine et de spécifier son vocabulaire de représentation (Gruber, 1993). Dans notre cas, le fait d'associer des éléments qui composent le réseau à une ontologie nous permet d'organiser les objets qui nous semblent pertinents pour la construction du réseau du *Maitron*. L'ontologie est une organisation qui se prête aussi à l'évolution de notre compréhension de ces objets car elle nous permet de prévoir des niveaux de granularité. Finalement, cette manière de décrire les objets qui composent notre représentation est au cœur de sa portabilité (voir section 5).

Dans un premier temps, nous nous intéressons principalement au niveau le plus générique d'interaction entre les objets du réseau. Le point de départ, inspiré de l'ARS, sont les individus et leur relations. Dans le graphe qui représente le réseau, les individus correspondent donc aux sommets, et les relations aux arcs. Nous considérons cependant que le réseau du *Maitron* est structuré en bonne partie par l'interaction entre individus et d'autres types d'entités.

Au premier niveau de notre ontologie, nous plaçons donc deux types d'objets, les RELATIONS, et les ENTITÉS. Nous distinguons quatre types d'entités : INDIVIDUS, ORGANISATIONS, PUBLICATIONS et ÉVÉNEMENTS. Quant aux relations, elles sont définies en fonction des différents rapports qu'entretiennent deux entités. Chaque entité est susceptible d'être en relation avec une autre : les individus peuvent être en relation avec d'autres individus, mais aussi avec des organisations, des publications ou des événements. Dans la figure 1 nous trouvons la hiérarchie d'objets (à gauche) et la combinatoire des relations (à droite).

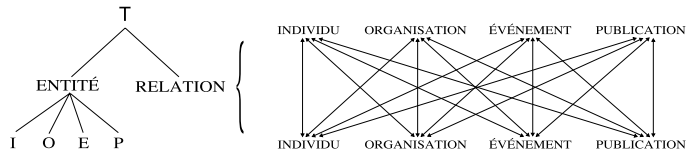


FIG. 1 – Ontologie des objets et combinatoire de relations entre entités

En langue naturel ces connections peuvent prendre des formes différentes. Par exemple, si l'on pense aux liens possibles entre INDIVIDUS et les quatre autres types d'ENTITÉS, on trouvera rapidement des marqueurs (des éléments de la langue) qui expriment ces relations : ainsi, on dira

que les individus *rencontrent* des individus, *participent* à des événements, sont *membres* d'organisations, *écrivent* pour des publications. Pour chaque type de relation il y a de nombreuses manières d'exprimer les liens entre les deux types d'entités concernés.

Pour construire le réseau de ces relations, les structures linguistiques qui sont utilisées pour les exprimer doivent être identifiées, modélisées et enfin trouvées dans le corpus. En retrouvant les instances spécifiques on peut extraire les informations nécessaires pour décrire les liens. Nous utilisons des techniques d'EI sur du texte prétraité à l'aide d'outils de TAL pour faire cette extraction.

3.3 L'extraction d'entités et de relations

D'autres travaux d'EI se sont déjà intéressés à l'extraction automatique de relations et à la représentation d'informations biographiques extraites de manière automatique. Nous présenterons brièvement quelques uns.

Riloff (1996) propose un système capable de générer ce qu'il appelle des *case frames* (qui seraient grosso modo équivalents aux formulaires d'extraction) comportant plus de deux entités dans un scénario d'extraction. Le repérage des relations entre entités est utilisé dans ce système dans le but de construire automatiquement des formulaires d'extraction plus complexes. L'objectif de Riloff est de proposer un système qui s'adapte facilement à un nouveau domaine (et une nouvelle tâche) d'extraction. L'extraction même consiste, encore une fois, à retrouver l'information nécessaire pour remplir des formulaires isolés.

Le point de départ du système REES (Aone & Ramos-Santacruz, 2000) est deux ontologies. Une première ontologie est composée d'une typologie d'entités et les relations qu'on leur associe, la deuxième ontologie est une classification de types d'événements. Les relations sont définies entre les quatre types d'entités génériques et les attributs qu'on leur associe généralement. Pour un type d'entité comme « personne », par exemple, on définit des relations avec un titre, une nationalité, un numéro de téléphone, une affiliation, un type, un soustype, etc. Quant aux événements, ils sont composés d'un ensemble de participants (les entités). L'approche cherche à donner de la généricité au système d'EI par une définition *a priori* des événements possibles ; des événements qui sont présentés comme plus génériques. Il n'y a pas de liens entre les instances isolées d'événements qui sont extraites.

Le Priol (2001) utilise la méthode dite d'exploration contextuelle pour extraire automatiquement des relations sémantiques comme, par exemple, *est partie de* ou est *inclu*. A la différence des systèmes d'extraction classiques, le système de Le Priol se concentre sur les liens sémantiques entre unités lexicales, et ne s'intéresse pas aux événements. Du fait de mettre des unités lexicales en relation, le système vise, comme nous, de produire une représentation complexe. En revanche, cette représentation est plus proche d'une ontologie extraite automatiquement pour représenter des connaissances génériques que d'un réseau social qui représente des informations liées à des événements et des acteurs.

Shinyama and Sekine (2006) utilisent des techniques de *clustering* statistique pour découvrir automatiquement des relations dans un corpus de dépêches de presse. Leur objectif est de constituer l'ensemble de relations qui peuvent être utilisées plus tard pour construire des formulaires d'extraction. Les formulaires résultent des régularités dans les structures linguistiques du corpus. Cette approche pourrait être utilisée dans le cadre de notre travail. En effet, les ingénieurs linguistes qui développent des patrons d'extraction pourraient utiliser le *clustering* statistique

pour découvrir des structures récurrentes.

Finalement, Kevers (2006) s'intéresse à l'extraction automatique de données biographiques et à leur représentation. Son objectif est de décrire un modèle de représentation d'information biographique fondé sur des triplets (*sujet, relation, objet*). Son modèle s'applique donc à la description des attributs biographiques que l'on a mentionnés dans la section 3.1 comme étant secondaires à la structuration du réseau de relations entre entités. Son travail peut donc être vu comme étant complémentaire au nôtre.

4 Extraction de structures linguistiques

Maintenant que nous avons une meilleure compréhension de la structure de la représentation que l'on veut produire et des objets qui la composent, nous devons décrire les techniques utilisées pour récolter et arranger ces objets dans cette structure.

On distinguera deux étapes dans ce processus. La première consiste à analyser le texte non structuré, c'est-à-dire à l'enrichir avec des annotations linguistiques. La deuxième étape est l'extraction elle-même, qui opère sur les structures qui résultent de la première étape.

4.1 Annotation linguistique du corpus

Nous utilisons des outils de l'architecture de développement linguistique MACAON[3] pour le prétraitement des textes du corpus. Ce prétraitement est indépendant de l'extraction et pourrait se faire avec des outils différents. C'est pour cette raison que nous n'entrons pas ici dans les détails techniques de ces outils de TAL, et nous nous limitons à la description des caractéristiques pertinentes pour la suite.

Les outils MACAON sont une collection de modules de TAL dévéloppés pour des tâches spécifiques d'annotation de textes. Avant d'utiliser les patrons d'extraction associés aux différents types de relations entre entités, le texte des biographies est analysé avec des modules MACAON qui s'occupent des tâches suivantes :

1. Segmentation en phrases
2. Tokenisation
3. Reconnaissance d'entités nommées
4. Analyse lexicale
5. Étiquettage morpho-syntaxique
6. Analyse morphologique
7. Analyse syntaxique partielle

La reconnaissance d'entités nommées consiste à repérer des formes superficielles d'objets de type ENTITÉ de l'ontologie présentée dans la section 3.2.

L'analyseur partiel suit le modèle proposé par Abney (1996) mais incorpore aussi le concept de tête. Dans le modèle d'analyse partiel de Abney, des grammaires régulières décrivant des

[3] `http://code.google.com/p/macaon/`

constituants partiels (ou *chunks*) s'appliquent en cascade aux séquences de parties du discours qui composent les phrases du texte à analyser. Dans l'implémentation de MACAON, les patrons réguliers qui décrivent les *chunks* peuvent identifier l'un de ses composants comme étant la tête de l'ensemble. Cette application, inspirée de certains formalismes linguistiques (eg. HPSG), permet au module d'extraction d'accéder plus aisément à certains éléments de la phrase.

Après le passage par cet ensemble de modules de TAL, les phrases du corpus sont interprétées[4], au moment de l'extraction, comme des séquences de structures de traits. Les structures de traits contiennent l'ensemble des annotations linguistiques ajoutées par les différents modules.

4.2 FSMs et structures de traits

La reconnaissance des fragments de texte associés aux liens qui composent le réseau, ainsi que l'extraction des arguments dans chaque instance d'une relation sont tous les deux accomplis avec des machines à états finis (FSMs) à base de structures de traits (ou FS-FSMs[5]).

Nous venons d'expliquer que la version analysée des phrases du corpus est composée de structures de traits. Le module d'extraction utilise des machines à états finis pour reconnaître des patrons non pas sur un alphabet de symboles mais sur un alphabet de structures de traits. A la différence des automates traditionnels de reconnaissance, les FS-FSMs n'utilisent pas la relation de l'égalité lors de la reconnaissance. Cette relation entre les symboles de l'entrée de l'automate est remplacée par la relation de subsomption entre structures de traits.

Dans une machine à états finis habituelle, la reconnaissance se fait au fur et à mesure que la machine change d'état, et ceci en fonction des symboles qu'elle trouve sur la bande d'entrée et des symboles associés à ses transitions. Pour changer d'état, une FS-FSM qui se trouve dans un état donné doit avoir une transition associée à une structure de traits qui subsume la structure de traits de la bande d'entrée. Parallèlement, les entités sont extraites durant la reconnaissance. Ce n'est que lorsque la FS-FSM atteint un état d'acceptation (après avoir traversé un certain nombre de transitions) que l'instance de relation est reconnue et que l'information extraite est validée. Les instances incomplètes sont rejetées.

Une machine d'extraction simple décrit un patron spécifique utilisé pour la reconnaissance d'une relation. Ce patron pourrait être, par exemple, le patron décrivant des phrases avec le nom d'une personne (une entité de type INDIVIDU) suivi de « était membre de », suivi à son tour d'une entité nommée reconnue comme le nom d'un parti politique (une entité de type ORGANISATION). Des machines plus complexes peuvent être définies pour des patrons plus expressifs ; par exemple, pour l'ensemble de patrons construits autour du déclencheur lexical *membre*.

L'utilisation d'une technique de traitement derivée des machines à états finis nous permet de tenir compte de leurs propriétés et d'utiliser les opérations définies sur elles en tant que modèle de calcul. Nous pensons plus précisément à l'union, la minimisation et la déterminisation. Le module d'extraction peut être vu comme l'union des FS-FSMs associées aux relations entre les différents types d'entités. La minimisation et la déterminisation pourraient[6] être utilisées pour

[4]Voir (Gazdar *et al.*, 1988) pour une description de l'utilisation des structures de traits pour représenter des catégories syntaxiques.

[5]On préfère l'acronyme de la version anglaise « Feature Structure Finite State Machines » du fait de l'usage courant dans la littérature de l'acronyme FSM pour les machines à états finis

[6]Nous n'avons pas encore formalisé l'application des stratégies de minimisation ni de déterminisation à des

améliorer la performance (en temps et en espace) du module d'extraction.

Par ailleurs, le processus de la construction de la représentation peut aussi être vu comme un processus incrémental. Nous n'avons pas besoin de définir toutes les relations (et tous les patrons associés à ces relations) au préalable et la totalité du réseau ne doit pas être la sortie d'une seule extraction. L'extraction, telle qu'elle est décrite ici, peut opérer comme une méthode permettant l'acquisition incrémentale de connaissances. La taille de l'ensemble de relations augmente, ainsi que la taille du réseau, au fur et à mesure que les patrons des FS-FSMs sont affinés.

A l'heure actuelle, nous nous concentrons sur la mise au point des outils décrits dans la section 4.1 et sur la modélisations des patrons d'extraction, ce qui nous permettra de passer à l'étape d'évaluation des résultats. Bien que l'on ne puisse pas présenter pour l'instant des résultats concrets, nous pouvons mentionner les perspectives qui guident notre travail.

5 Perspectives

Nous voyons, au moins, trois directions d'évolution du travail que nous menons sur le *Maitron* : l'étude du réseau qui résulte de l'extraction, l'adaptation de la structure de représentation à d'autres domaines d'application et l'inclusion de connaissances extraites dans des systèmes intelligents.

Nous avons mentionné que l'ARS attire l'attention de plusieurs domaines de recherche. On pourrait adapter les avancées dans d'autres domaines à notre utilisation du corpus. L'application de méthodes de théorie des graphes, par exemple, pourrait aider à exploiter la structure du réseau. Avec les méthodes d'analyse de cette théorie, on peut envisager de retrouver les entités les plus saillantes, ou celles qui créent des ponts entre différents fragments du réseau. Nous pouvons aussi exploiter la structure du réseau pour raffiner la classification des entités du domaine dans le but, par exemple, d'adapter l'exploitation du réseau à la construction automatique d'ontologies.

Nous nous intéressons dans notre travail aux relations entre différents types d'entités. Bien que ces relations semblent spécifiques à l'interprétation de la structure du corpus du *Maitron*, elle ne sont pas limitées au domaine des corpus biographiques. Tant que les mêmes types d'entités sont concernées, les mêmes relations et les mêmes patrons peuvent être utilisés pour des tâches d'extraction et d'acquisition de connaissances dans d'autres domaines. L'Internet et ses communautés virtuelles, par exemple, sont un vaste domaine dans lequel on trouve les mêmes types d'entités. Les textes (ou autres données) partagés par les membres de ces groupes peuvent être utilisés pour constituer un corpus qui permettrait de recréer le réseau de leurs relations.

Nous avons choisi de représenter le réseau avec des déclarations RDF[7]. Ces dernières peuvent être intégrées directement dans les systèmes experts qui sont élaborés suivant les standards XML proposés pour la représentation et l'exploitation des données en ligne. De plus, ces standards entretiennent un lien étroit avec le domaine de la représentation de connaissances ; notons que notre modélisation du réseau peut être traduite en termes de logique de description. Le réseau du *Maitron* structure des connaissances très précises, mais ces connaissances peuvent s'intégrer dans des systèmes qui exploitent des bases de connaissances pour d'autres domaines

machines qui utilisent la subsomption comme relation de reconnaissance. Leur union, en revanche, est triviale.

[7]La présentation de ces standard XML étant au-délà de la portée de cette article, nous dirigeons le lecteur vers la documentation qui peut se trouver en ligne.

d'application comme, par exemple, les systèmes de question-réponse.

6 Conclusion

Nous avons proposé d'utiliser des techniques d'EI dans le but de construire des structures de représentation plus complexes que les bases de données plates qui sont souvent utilisées pour l'évaluation des systèmes d'extraction.

Nous avons décrit une représentation organisée autour des relations entre types d'entités génériques. Le cas du travail en cours avec un corpus de biographies a été utilisé pour décrire une application concrète de cette manière d'exploiter les résultats d'un processus d'extraction dans le but de constituer une base de connaissances avec une structure inspirée de l'ARS.

Nous avons décrit la méthodologie et les outils utilisés dans le processus d'extraction, et nous nous sommes particulièrement intéressé à la description des machines à états finis qui modélisent des structures du langage naturel.

Finalement, nous avons présenté trois chemins distincts d'évolution de ce projet de recherche.

Références

ABNEY S. (1996). Partial Parsing via Finite-State Cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, p. 8–15, Prague.

AONE C. & RAMOS-SANTACRUZ M. (2000). REES: A Large-scale Relation and Event Extraction System. In *Proceedings of the 6th Applied Natural Language Processing Conference*.

GAZDAR G., PULLUM G. K., CARPENTER R., KLEIN E., HUKARI T. E. & LEVINE R. D. (1988). Category Structures. *Computational Linguistics*, **14**(1), 1–19.

GRUBER T. R. (1993). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. GUARINO & R. POLI, Eds., *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands: Kluwer Academic Publishers.

KEVERS L. (2006). L'information biographique: modélisation, organisation et extraction en base de connaissances. In *Actes de TALN-RECITAL*, p. 680–689, Leuven, Swisse.

LE PRIOL F. (2001). Identification, interprétation et représentation de relations sémantiques entre concepts. In *Actes de TALN*.

M. T. PAZIENZA, Ed. (1999). *Information Extraction: Towards Scalable, Adaptable Systems*, volume 1714 of *Lecture Notes in Artificial Intelligence*. Berlin, Germany: Springer.

POIBEAU T. (2003). *Extraction automatique d'information : du texte brut au web sémantique*. Paris, France: Hermès.

RILOFF E. (1996). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, p. 1044–1049.

SHINYAMA Y. & SEKINE S. (2006). Preemptive Information Extraction using Unrestricted Relation Discovery. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, p. 304–311, New York City, USA: Association for Computational Linguistics.

Vers une ressource prédicative pour l'extraction d'information

Aurélien BOSSARD
LIPN – Université Paris 13, 93200 Villetaneuse
aurelien.bossard@lipn.univ-paris13.fr

Résumé. Cet article présente une méthode pour construire, à partir d'une ressource lexicale prédicative existante, une ressource enrichie pouvant servir à une tâche d'extraction. Nous montrons les points forts et les lacunes de deux ressources existantes pour le Français : les Tables du LADL et Volem. Après avoir montré pourquoi nous avons sélectionné Volem, nous listons les données nécessaires à la tâche d'extraction d'information. Nous présentons le processus d'enrichissement de la ressource initiale et une évaluation, à travers une tâche d'extraction d'information concernant des textes de rachats d'entreprise.

Abstract. In this article, we present a method aiming at building a resource for an information extraction task, from an already existing French predicative lexical resource. We point out the weaknesses and strengthnesses of two predicative resources we worked with : Les tables du LADL and Volem. We present why we select Volem as the most interesting resource for the task. Thereafter, we make a list of the needs an information extraction task implies, and how we include missing information in the resource we selected. We evaluate the resource completed by those missing informations, using it in an information extraction task.

Mots-clés : ressource prédicative, extraction d'information, patrons lexico-syntaxiques.

Keywords: predicative resource, information extraction, lexico-syntactic patterns.

1 Introduction

Cet article vise à montrer la façon dont nous avons procédé pour contribuer à une ressource lexicale pour le français, utile à des fins d'extraction d'information, en nous servant de ressources déjà existantes. La création d'une ressource pour le français assez complète pour couvrir de larges champs d'application en TAL apparaît aujourd'hui comme un enjeu majeur. En effet, en raison du manque de ressources, seules deux approches sont aujourd'hui possibles afin de réaliser des applications du type extraction d'information : l'apprentissage automatique et/ou la création de règles à la main.

L'approche que nous défendons ici est fondée sur la notion de schéma prédicatif. Les éléments pertinents pour l'extraction sont ceux qui se situent autour d'une relation sémantique, portée par un nom prédicatif ou, plus souvent, par un verbe. L'étude des schémas prédicatifs permet d'attribuer un rôle à chacun des arguments du prédicat ; l'étude de la syntaxe de la phrase permet en outre une mise en relation des arguments d'un prédicat avec leur rôle thématique. Nous testons notre approche dans le cadre d'une application d'extraction d'information portant sur des

rachats d'entreprise. La tâche consiste par exemple à donner une représentation (sémantique) identique pour les trois phrases suivantes :

– CPI rachète Fulmar.
– CPI a racheté Fulmar à son PDG pour 50 millions d'euros.
– CPI a indiqué avoir racheté Fulmar.

Dans chacune de ces trois phrases, nous pouvons identifier un acheteur : *CPI*, une entreprise acheté : *Fulmar*, auxquelles peuvent s'ajouter des données sur le montant de la transaction, le vendeur, la date de la transaction... Ce type d'applications a déjà été développé, y compris pour le français (Poibeau, 2003). Notre but est ici de l'envisager avec un nouveau regard, en nous focalisant sur des ressources sémantiquement riches. Plutôt que de développer des ressources de manière *ad hoc*, nous cherchons à caractériser l'intérêt des données déjà existantes pour le français.

Les questions que nous nous sommes posé sont les suivantes :
– Quelles sont les informations qu'une ressource lexicale syntaxique doit encoder pour que l'on puisse arriver à un tel résultat ?
– Est-il possible, avec les ressources existantes, de créer une telle ressource ?
– Quel réel intérêt aurait une telle ressource (précision de l'extraction, rappel, automatisation...) ?

Dans un premier temps, nous présentons un état de l'art traitant des ressources lexicales syntaxique existantes. Dans une seconde partie, nous caractérisons plus en détail notre sujet d'étude avant d'aborder, dans une troisième partie, les expériences réalisées. Nous présentons ensuite les résultats et leur analyse avant de poser, dans une dernière partie, les conclusions de notre recherche.

2 Choix d'une ressource

L'anglais dispose aujourd'hui de trois ressources à large couverture encodant d'une manière ou d'une autre la notion de schéma prédicatif : VerbNet, PropBank et FrameNet. Ces trois ressources sont fondées sur des approches différentes : approche syntaxique pour VerbNet et PropBank, et approche sémantique pour FrameNet (Pitel, 2006). De nombreuses applications ont été développées autour de FrameNet, comme la désambiguïsation sémantique (Fillmore & Baker, 2001), (Lowe *et al.*, 1997), mais aussi l'extraction d'information. Des recherches ont été menées pour utiliser conjointement ces trois ressources afin d'améliorer l'étiquetage sémantique (Giuglea & Moschitti, 2004).

Il existe beaucoup moins de richesse pour le français. Le Dictionnaire Explicatif et Combinatoire (DEC, `http://www.olst.umontreal.ca/decfr.html`) d'I. Melc'ŭk a été exclu car il n'offrait pas une couverture suffisante pour la tâche. DicoValence (`http://bach.arts.kuleuven.be/dicovalence/`) n'était quant à lui pas disponible au moment de l'étude mais mériterait sinon d'être pris en considération. Nous nous sommes alors focalisés sur deux ressources pour le français : Volem et les tables du LADL. Il s'agit dans cette partie d'expliquer le choix que nous avons fait concernant la ressource que nous avons utilisée.

2.1 Les Tables du LADL

Les Tables du LADL, aussi connues sous le nom de lexique-grammaire, ont été établies sous la direction de Maurice Gross. Elles regroupent 6000 verbes répartis dans des tables construites d'après des similitudes de comportement syntaxique. Chaque table du Lexique-Grammaire contient un certain nombre de propriétés, qui sont validées ou invalidées pour chacun des verbes qui y figure (matrice de + et de −). Les propriétés encodent des informations sur (Gross, 1975) :
– Les réalisations possibles des arguments (restrictions de sélection : arguments à trait « humain » ou « non-humain », argument de type abstrait, objet...) ;
– Les propriétés syntaxiques du verbe ou de ses arguments (pronominalisation des verbes, introduction d'un argument par une préposition...)
– Les sous-catégorisations alternatives ;
– Les possibilités de redistributions (passif long, passif court...).
Les informations contenues dans les tables du LADL sont riches sur le plan syntaxique mais relativement pauvres sur le plan sémantique. Les arguments ne sont typés que par des restrictions de sélection (`humain`, `non-humain`, `objet concret`, `abstrait`...). Le format en colonnes des tables et le fait que l'information soit répartie sur plusieurs colonnes rend les traitements difficiles. Le fait que, selon les tables, les propriétés codées dans les colonnes ne sont pas toujours les mêmes complique encore le traitement. Il est nécessaire d'effectuer un travail important de transformation pour rendre ces tables exploitables directement par des applications de TAL (Gardent *et al.*, 2005).

2.2 Volem

Volem (Saint-Dizier *et al.*, 2002) est une ressource multilingue (français-espagnol-catalan). Les entrées sont des verbes : la ressource décrit leur comportement syntaxique et sémantique à travers la description des arguments et des schémas de sous-catégorisation. Cette ressource décrit à l'heure actuelle 1700 verbes.

Description du verbe : acheter

GRILLE THEMATIQUE :
[[inic(agent),dest],[th],[src]]
LCS :
ALTERNANCES :
caus_2np_pp , anti_pr_np , anti_pr_np_pp , pas_etre_part_np_2pp , pas_etre_part_np_pp , caus_2np , caus_refl_pr_2np , caus_np_pp , caus_support_np
WN :
[13,2,3], [13,3,1] , [13,3,8]
EXEMPLE :
Il a acheté ce livre à un brocanteur

FIG. 1 – L'entrée lexicale du verbe « acheter » dans Volem

Cette ressource est fondée sur une liste de rôles thématiques génériques, mais assez précis. Les différents rôles thématiques peuvent être combinés afin de décrire aux mieux les arguments d'un verbe (*cf.* Figure 1).

Les principaux inconvénients de cette ressource sont :

– L'absence de gestion de la polysémie (les concepteurs de la ressource ont fait le choix de ne coder qu'un sens par verbe, correspondant à l'emploi le plus fréquent) ;
– La faible couverture de la ressource (1700 verbes) ;
– L'absence de description précise des schémas syntaxiques que représentent les différentes alternances utilisées dans Volem.

Volem a une couverture moindre que celle des tables du LADL. L'absence de gestion des rôles thématiques dans les tables du LADL constitue cependant un inconvénient de taille pour une tâche d'extraction d'information, qui demande plus poussée sur la nature des arguments. Nous avons donc choisi de concentrer notre étude sur la ressource lexicale Volem, qui paraît avoir un potentiel de description plus fort que les Tables du LADL pour notre application d'extraction.

3 Méthode d'enrichissement de la ressource

Les informations contenues dans Volem n'étaient pas suffisantes pour réaliser une tâche d'extraction d'information. Le format de Volem au format XML n'était pas non plus directement exploitable. Nous avons donc retravaillé la ressource sur les points suivants afin de pouvoir l'utiliser dans le cadre d'une tâche d'extraction d'information :

1. Codage de la ressource sous la forme d'une table de contraintes
2. Ajout des informations manquantes
3. Codage d'automates patrons

3.1 Codage de la ressource sous la forme d'une table de contraintes

Nous voulons, à partir de Volem, créer des automates d'extraction (format unitex : `http://www-igm.univ-mlv.fr/~unitex/`). Pour cela, nous avons besoin d'une ressource codée à l'aide d'un tableau. Nous avons donc créé un convertisseur pour passer du format de Volem à notre format. Nous utilisons une colonne par alternance de Volem, et validons l'alternance pour un verbe en mettant un « + » dans la case correspondante, et un « - » sinon. Pour encoder les informations sur les rôles thématiques, nous écrivons une colonne par argument du verbe (3 colonnes) et nous remplissons chacune d'elle avec la description enregistrée au sein de Volem du rôle thématique de l'argument. Ainsi, la ressource est directement exploitable par des graphes unitex, qui exploitent les tables de contraintes.

3.2 Ajout des informations manquantes

Cependant, la ressource en l'état n'encode toujours pas assez d'informations pour réaliser une extraction d'information précise. En effet, il lui manque encore :

1. les auxiliaires des verbes
2. les différentes prépositions introduisant éventuellement un argument
3. les noms associés aux verbes (*e.g.* *rachat* pour *racheter*)
4. les adjonctions essentielles à une relation à extraire.

3.2.1 Les auxiliaires

Deux possibilités se sont offertes à nous pour ajouter les auxiliaires d'un verbe à la table de données que nous avions construite. Soit récupérer les auxiliaires depuis un dictionnaire, soit identifier l'auxiliaire d'un verbe grâce aux occurrences de celui-ci en corpus. Nous avons opté pour la deuxième solution. A partir de la table de données que nous avons créée, un automate est créé qui reconnaît les différents verbes ainsi que les auxiliaires qui les accompagnent. Si l'auxiliaire « avoir » apparaît au moins une fois dans le texte pour un verbe donné, un « + » est ajouté à l'intersection de la ligne correspondant à ce verbe et de la colonne correspondant à l'auxiliaire « avoir ». L'inconvénient de cette méthode est qu'elle demande des textes correctement écrits. Mais elle n'est pas dépendante d'une ressource comme l'aurait été la première : en effet, en utilisant les tables du LADL comme référence, nous n'aurions pas pu ajouter automatiquement certains verbes qui ne sont pas encodés dans la ressource. Nous avons procédé au repérage des auxiliaires avec la version « brute » des corpus que nous avons utilisés pour extraire les relations de rachat d'entreprises (*cf.* §4.1).

3.2.2 L'ajout des prépositions

Nous avons déjà mentionné qu'une seule préposition est codée par argument au sein de Volem, alors que plusieurs prépositions peuvent apparaître pour certains verbes (*acheter à* ou *auprès de*). Nous avons donc réalisé un outil permettant d'ajouter à la table de données les prépositions qui introduisent les arguments d'un verbe. Cet outil nécessite une validation des résultats par l'utilisateur.

Le système est fondé sur une série d'automates « à trou » : chaque « trou » correspond à une préposition possible introduisant un argument. Le système renvoie quelques erreurs (soit des groupes de mots qui ne sont pas des prépositions, soit des prépositions rallongées de certains mots qui les suivaient dans le texte). Après validation des résultats par l'utilisateur, les prépositions validées sont ajoutées à la table de données. Cet outil nécessite un corpus annoté par entités nommées. Nous avons travaillé sur les corpus utilisés pour l'extraction de relations de rachats d'entreprise (*cf.* §4.1).

3.2.3 Les adjonctions

Volem ne gère que les arguments clé d'un verbe. Cependant, certains arguments qui ne sont pas nécessaires d'un point de vue syntaxique jouent un rôle extrêmement important dans des relations à extraire. Par exemple, un achat selon Volem ne fait pas intervenir de montant : le montant est quasiment toujours présent quand on se base sur l'analyse en corpus. C'est donc un argument clé, sémantiquement parlant.

L'approche développée par les auteurs de FrameNet est du même type (Fillmore & Baker, 2001). La description des schémas prédicatifs au sein de cette ressource se fonde sur l'étude en corpus des réalisation du verbe. Si un complément intervient fréquemment pour un verbe donné, alors celui-ci sera assimilé à un argument, même s'il est considéré comme un ajout dans la grammaire traditionnelle.

Nous avons alors tenté, en dénombrant ces adjonctions au sein des corpus étudiés, de déterminer quelles adjonctions essentielles pouvaient tenir lieu d'argument et dans quelle mesure celles-ci

pouvaient être repérées par une analyse statistique. La méthode se fonde sur le repérage et le regroupement des compléments circonstanciels (temps, lieu, montant...) pour chaque verbe au sein du corpus.

Après dénombrement des adjonctions, nous ne retenons que celles au-dessus d'un seuil de 10 % (défini manuellement). Cela signifie que nous ne sélectionnons que celles qui sont apparues dans au moins 10 % des phrases contenant un verbe donné. Cette méthode nous a permis de compléter nos informations concernant les données à extraire avec les adjonctions adéquates pour 80 % des verbes sélectionnés (pour les 20 % restant, l'adjonction « montant » était apparue dans moins de 10 % des phrases à extraire, contre 15 % pour l'adjonction « date »). Les résultats doivent toutefois être validés par un expert avant d'ajouter les adjonctions au schéma argumental des verbes concernés.

3.3 Les automates patrons

La dernière étape de l'enrichissement de la ressource a consisté en la création d'automates patrons pour chacune des alternances listée dans Volem. Un automate patron est un automate lexicalement vide, encodant une famille d'alternances ; il est instancié par l'ensemble des verbes correspondant à la famille d'alternances visée. Pour cela, il a fallu dans un premier temps identifier les différentes formes de surface que présentent chacune des alternances de Volem, puis réaliser pour chacune d'entre elles des graphes qui permettent de les reconnaître.

La figure 2 (*cf.* dernière page de l'article) présente un extrait d'un graphe qui permet de reconnaître l'alternance *caus_2np* de Volem.

Ces automates patrons prennent en entrée la ressource que nous avons créée, et produisent en sortie autant d'automates que d'alternances à reconnaître pour chaque verbe. Les automates ainsi créés annotent un texte avec les informations que l'on veut extraire. En l'occurrence, pour notre extraction sur les rachats d'entreprise, ils reconnaissent les fragments de textes correspondant à l'acheteur, au vendeur, à l'élément vendu, au montant et à la date.

4 Expériences

4.1 Données d'évaluation

Nous avons choisi de travailler sur une tâche d'extraction précise : le rachat d'entreprises. Les expériences ont été menées sur plusieurs corpus :
– Un corpus tiré d'un autre site spécialisé : FUSACQ (300ko, 25000 mots) (`http://www.fusacq.com`)
– Un corpus tiré de différents journaux généralistes (400ko)
– Un corpus d'entraînement pour repérer les entités nommées (200ko).

L'utilisation de plusieurs corpus permet d'évaluer les performances en tenant compte (dans la mesure du possible) du genre textuel. On ne trouve pas les mêmes constructions ni les mêmes expressions suivant que l'on a affaire à un corpus journalistique ou à un site web. Nous verrons dans la discussion que cette hypothèse se vérifie lors de l'étude des performances sur les différents corpus.

4.2 Résultats

Nous avons mis en place deux protocoles d'évaluation, afin d'isoler les éventuels problèmes ; dans l'un, nous passons les règles d'extraction sur un corpus dans lequel les entités nommées ont été annotées grâce à un outil pour annoter développé au LIPN (TagEN, `http://www-lipn.univ-paris13.fr/~poibeau/tagen.html`). Dans l'autre, nous utilisons un corpus dans lequel nous avons annoté toutes les entités nommées à la main. Nous pouvons ainsi procéder d'un côté à une évaluation « en conditions réelles », et de l'autre, de nous focaliser sur l'évaluation des schémas prédicatifs, indépendamment des erreurs dues à la mauvaise reconnaissance des entités.

Dans le tableau 1, nous entendons « relations » comme des structures grammaticales comportant un verbe et ses arguments participant à un rachat d'entreprise.

	Protocole 1	Protocole 2	Nombre total de relations
Nombre de relations repérées	101	184	285
% de relations	35	64	100

TAB. 1 – Tableau des résultats de l'extraction sur le corpus FUSACQ

Un peu plus seulement de la moitié des entités nommées correspondant à un acheteur potentiel ou à un vendeur potentiel ont été annotées. L'annotation des entités nommées n'a pas été menée plus avant, étant donnée qu'elle n'est pas au centre de notre étude. Notre outil permet de repérer dans un texte dans lequel l'annotation des entités nommées est correcte, 65 % des relations d'achat.

35 % des relations restent tout de même non repérées. Ceci provient du fait que nous n'avons pas géré certains schémas syntaxiques :
– les subordonnées relatives
– les verbes introducteurs précédés d'un verbe marquant soit le passé, soit le futur (ambiguïté sémantique possible. Ex. : « Bull vient d'annoncer le rachat de CP8 à Schlumberger »)
– les structures complexes (ex. :« COMPANY1 s'était diversifié à travers l'acquisition de COMPANY2 »
– L'alternance passive sans groupe prépositionnel (non encodé dans Volem pour les verbes qui nous intéressent. Ex. : « COMPANY a été racheté »)
– Les structures faisant intervenir un pronom (pas de résolution d'anaphores).
L'outil de repérage des prépositions renvoie des résultats bruités à hauteur de 8 %.

Les adjonctions nécessaires à une tâche d'extraction sont la date et le montant de la transaction. Les corpus sur lesquels nous avons fait nos expériences ont montré ce point, même s'ils sont de taille trop faible pour donner des chiffres significatifs statistiquement.

4.3 Discussion

Nous avons vu que Volem est incomplet du fait qu'il ne gère pas la polysémie et que toutes les alternances n'y sont pas codées.

Est-il possible de rajouter aux entrées de Volem les alternances que cette ressource n'encode

Verbe	Alternances	nombre d'occurrences de l'alternance (FUSACQ)	nombre d'occurrences de l'alternance (Corpus général)
racheter	caus_2np	37	16
	caus_2np_pp	6	12
	pas_etre_part_np_pp	6	12
revendre	caus_2np	1	0
acheter	caus_refl_pr_np	2	0
	caus_2np	0	16
vendre	pas_etre_part_np_2pp	2	0
	caus_2np	2	0
	caus_2np_pp	2	0
acquérir	caus_2np	44	0
	caus_2np_pp	1	4
	pas_etre_part_np_pp	0	16
céder	caus_2np_pp	24	4
	caus_2np	9	0
	pas_etre_part_np_2pp	2	8
fusionner	aucune occurrence	0	0
détenir	caus_2np	5	0
	pas_etre_part_np_pp	1	12
offrir	caus_2np_pp	1	16
	caus_2np	1	0
reprendre	pas_etre_part_np_pp	11	16
	caus_2np	12	8

TAB. 2 – Répartition des alternances selon les verbes dans les phrases extraites des corpus (FUSACQ annoté et extrait du corpus général)

pas ? L'ajout des alternances constitue un réel problème. En effet, il est possible, par des méthodes statistiques, de sélectionner des schémas de sous-catégorisation acceptables pour un verbe (Salmon-Alt & Chesley, 2005),(Briscoe & Carroll, 1997). Mais la sélection d'alternances acceptables constitue un tout autre problème ; il faut en effet pouvoir distinguer une phrase du type : « COMPANY1 a acheté une usine à LIEU. » d'une phrase du type : « COMPANY1 a acheté des terrains à LIEU{la ville de LIEU}. Ce problème mériterait une étude approfondie.

Un point intéressant est la variation dans l'usage du verbe suivant le corpus. On s'aperçoit, même sur des corpus de taille modeste, des variations d'usage, une alternance étant plutôt employée dans un corpus, une autre dans un autre corpus (*cf.* tableau2). Nous faisons l'hypothèse qu'il s'agit de variations dans le style d'écriture propre aux différents genres textuels. Ainsi, l'alternance « caus_2np » du verbe « détenir » constitue 78 % des variations syntaxiques pour le verbe « détenir » dans le corpus FirstInvest, et 80 % dans FUSACQ, mais n'apparaît pas dans le corpus tiré de journaux non spécialisés.

Les résultats obtenus sont satisfaisants. En effet, l'ajout des structures syntaxiques manquantes (*cf.* §4.2) permettrait d'obtenir environ 65 % de rappel sur une tâche d'extraction des rachats d'entreprise, soit 15 % de moins que le rappel obtenu par Thierry Poibeau, sur la même tâche et le même type de corpus (Poibeau, 2003), mais en utilisant une méthode à base de ressources qui se distingue des autres travaux par une certaine généricité.

5 Conclusion et Perspectives

Les ressources pour le français sont beaucoup moins complètes que celles pour l'anglais. La ressource pour le français qui nous a semblé la plus adaptée à l'extraction d'information (Volem), présente des manques (non gestion de la polysémie, couverture faible, alternances non encodées...) qu'il est cependant possible de combler par des méthodes semi-automatiques.

Les résultats obtenus pour une tâche d'extraction pour l'anglais (Giuglea & Moschitti, 2004) montre que l'extraction à base de ressources à large couverture permet d'obtenir de bons résultats et évite de redévelopper de manière *ad hoc* des connaissances pour chaque nouvelle application.

Il reste certes un travail à réaliser dépendant de l'application (ajout d'adjonctions, prépositions), cependant la mise en place d'une méthode utilisant des ressources extérieures permet une réutilisabilité *a contrario* des méthodes purement *ad hoc* tout en garantissant malgré tout une couverture et une certaine généricité.

Cet article a cherché à montrer comment compléter Volem, notamment avec les prépositions et le filtrage des alternances non pertinentes pour un sens donné. Il reste à définir une méthode semi-automatique pour l'ajout des alternances non référencées par Volem.

Références

BRISCOE T. & CARROLL J. (1997). Automatic extraction of subcategorization from corpora.

FILLMORE C. & BAKER C. (2001). Frame semantics for text understanding. In *WordNet and Other Lexical Resources Workshop*, NAACL, Pittsburgh.

GARDENT C., GUILLAUME B., FALK I. & PERRIER G. (2005). Le lexique-grammaire de m. gross et le traitement automatique des langues.

GIUGLEA A.-M. & MOSCHITTI A. (2004). Knowledge discovering using framenet, verbnet and propbank. In *International Workshop on Mining for and from the Semantic Web*, Seattle, USA.

GROSS M. (1975). *Méthodes en syntaxe*. Paris : Hermann.

LOWE J., BAKER C. & FILLMORE C. (1997). A frame-semantic approach to semantic annotation.

PITEL G. (2006). Framenet, théorie, produit, processus, multilingualité et connexions. In *Autour de FrameNet et de la Sémantique Lexicale Multilingue : projets en cours et points de contacts entre les différentes approches*. date de la conférence : 28 Février 2006.

POIBEAU T. (2003). *Extraction automatique d'information, du texte brut au web sémantique*. Paris : Hermes.

SAINT-DIZIER P., FERNANDEZ A., VAZQUEZ G., KAMEL M. & BENAMARA F. (2002). The Volem Project : a Framework for the Construction of Advanced Multilingual Lexicons . In *Language Technology 2002 , Hyderabad,* , p. 123–142 : Springer Verlag, Lecture Notes. Dates de conférence : décembre 2002.

SALMON-ALT S. & CHESLEY P. (2005). Le filtrage probabiliste dans l'extraction automatique de cadres de sous-catégorisation. In *Journée ATTALA du 12/03/2005*.

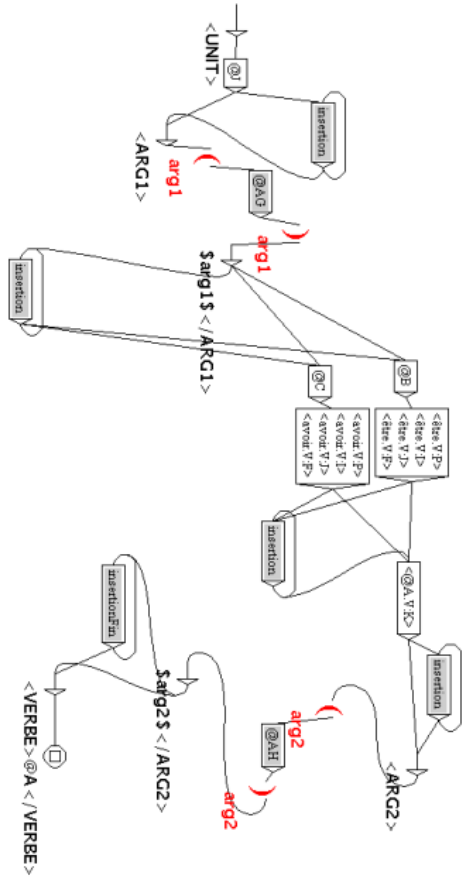


FIG. 2 – Exemple d'un automate patron

Caractérisation d'un corpus de requêtes d'assistance

François BOUCHET
LIMSI-CNRS, Université Paris-Sud XI, BP 133, 91403 Orsay Cedex
bouchet@limsi.fr

Résumé. Afin de concevoir un agent conversationnel logiciel capable d'assister des utilisateurs novices d'applications informatiques, nous avons été amenés à constituer un corpus spécifique de requêtes d'assistance en français, et à étudier ses caractéristiques. Nous montrons ici que les requêtes d'assistance se distinguent nettement de requêtes issues d'autres corpus disponibles dans des domaines proches. Nous mettons également en évidence le fait que ce corpus n'est pas homogène, mais contient au contraire plusieurs activités conversationnelles distinctes, dont l'assistance elle-même. Ces observations nous permettent de discuter de l'opportunité de considérer l'assistance comme un registre particulier de la langue générale.

Abstract. In order to conceive a conversational agent able to assist ordinary people using softwares, we have built up a specific corpus of assistance requests in french, and studied its characteristics. We show here that assistance requests can be clearly distinguished from the ones from other available corpora in related domains. We also show that this corpus isn't homogenous, but on the contrary reflects various conversational activities, among which the assistance itself. Those observations allow us to discuss about the opportunity to consider assistance as a general language particular registre.

Mots-clés : corpus de requêtes d'assistance, agent conversationnel, activité conversationnelle, actes de dialogue.

Keywords: corpus of assistance requests, conversational agent, conversational activity, speech acts.

1 Introduction

Le développement de l'informatique pour le grand public a entraîné une forte augmentation du nombre d'usagers novices en informatique en contact régulier avec celle-ci. Ces usagers novices n'ont bien souvent ni le temps ni l'envie d'utiliser des manuels papiers ou des FAQ (Foires Aux Questions) de logiciels de plus en plus complexes (en dépit des progrès ergonomiques) dont par ailleurs ils ne maîtrisent pas le vocabulaire spécifique. Des systèmes d'aides contextuelles (ou CHS en anglais, pour Contextual Help Systems (Jansen, 2005)) ont été développées pour mieux s'adapter aux besoins de ces nouveaux utilisateurs, mais ceux-ci semblent toujours préférer faire appel à un ami expert lorsqu'ils souhaitent réaliser une tâche particulière dans une application (Capobianco & Carbonell, 2002).
Parallèlement, les agents conversationnels animés dotés de capacités de dialogue et de raisonnement de niveaux variés (Sadek *et al.*, 1997) développés récemment ont mis en évidence les nom-

breux avantages potentiels d'une présence (même virtuelle) pour faciliter l'interaction homme-machine (Lester *et al.*, 1997).

Pour répondre à ce besoin d'assistance des usagers novices, le projet DAFT développé au LIMSI-CNRS (Sansonnet *et al.*, 2005) se propose donc de développer des Agents Conversationnels Assistants (ACA), capables d'analyser des requêtes en langue naturelle écrite non contrainte provenant d'usagers novices en situation réelle d'utilisation d'applications de complexités diverses (applets simples, pages web, traitement de texte). Pour répondre à ce type de requêtes, le système d'assistance raisonne sur la structure et le fonctionnement des applications à l'aide d'un modèle de celles-ci, construit de manière semi-automatique. Cette méthodologie a pour objectif de fournir une assistance pertinente en contexte à la manière des CHS, avec en plus tous les bénéfices liés à la présence d'agents conversationnels animés.

Afin d'identifier précisément les propriétés et les besoins propres à la Fonction d'Assistance, nous avons été amenés à constituer un corpus de requêtes, auquel on se référera sous le nom de corpus DAFT. Ce corpus illustre les actes de dialogues réalisés par des sujets en situation d'assistance et a été étudié lors d'un stage de M2R (Bouchet, 2006) afin de pouvoir réaliser les spécifications d'un langage de requêtes formelles adapté au domaine d'étude[1].
Dans cet article, nous détaillons dans un premier temps le processus de constitution du corpus employé, en justifiant en particulier la nécessité des choix effectués et en vérifiant que le corpus ainsi obtenu est viable pour répondre à nos besoins. Dans un second temps, nous nous proposons de comparer ce corpus avec d'autres corpus de domaines proches pour en dégager les spécificités. Enfin, nous analysons les différentes activités couvertes par les requêtes recueillies et tentons de les caractériser par des paramètres d'ordre linguistique.

2 Collecte et construction du corpus

2.1 Méthodologie employée

Au moment de l'étude, le corpus DAFT se composait d'environ 5 000 requêtes isolées (cf. §2.3) recueillies entre juin 2004 et juin 2006. Pour le constituer, nous avons eu recours à deux méthodes complémentaires garantissant à la fois l'empirisme et la bonne couverture du corpus :

1. Recueillir des requêtes réelles produites par des utilisateurs placés devant des applications intégrant un ACA de type LEA[2] (2/3 du corpus final).

2. Utiliser des structures dialogiques génériques issues de classifications de thésaurus anglais (600 structures (Molinsky & Bliss, 1994)) ou bilingues (300 structures (Atkins & Lewis, 1996)). Ces structures ont été adaptées pour les employer dans des requêtes d'assistance formulées dans le contexte des applications mentionnées dans le point précédent, afin d'assurer une certaine homogénéité lexicale du corpus (1/3 du corpus final). Ainsi, la structure "ça te dirait de…" utilisée conjointement avec une phrase recueillie comme "peux-tu jouer à ma place ?" donne "ça te dirait de jouer à ma place ?".

[1] Une première version moins complète de cette étude a été présentée lors de l'atelier WACA'02, n'ayant pas donné lieu à une parution d'actes.

[2] LIMSI Embodied Agent, développé par J-C. Martin dans le cadre du projet NICE (Buisine & Martin, 2005)

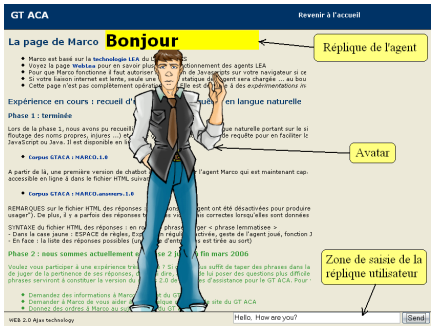


FIG. 1 – L'agent WebLea (Marco) placé sur une page du site web du GT ACA

L'utilisation de la seconde méthode nous permet de compenser la difficulté à obtenir un corpus de taille supérieure avec le panel réduit de sujets dont nous disposions[3]. Il y a recouvrement partiel avec les phrases recueillies par la première méthode, mais ce choix, discutable dans l'absolu, doit être mis en relation avec l'objectif de ce corpus qui est de fournir une base pour la constitution d'un agent rationnel : la couverture prime donc sur l'exactitude de la fréquence des phénomènes linguistiques. Le corpus ainsi construit sature donc mieux le domaine d'étude, même si cela induit fatalement un biais en sur-représentant des phénomènes linguistiques rares (les structures employées ne seraient apparues "naturellement" que sur un corpus plus grand).

2.2 Recueil des requêtes d'assistance

Les phrases recueillies l'ont été au sein de deux types d'applications :

1. deux applications de type Java (Le Guern, 2004) : un simple compteur temps réel (thread Java) dont l'utilisateur contrôle le démarrage et la vitesse, et un jeu de tours de Hanoi fonctionnant de manière modale (ie n'évoluant que si l'utilisateur agit).

2. deux sites web : une version active du site du groupe AMI du LIMSI permettant l'édition de contenu, et le site du GT ACA[4](cf. figure 1) en libre accès sur internet aux utilisateurs effectifs du site.

Dans la mesure où l'agent assistant que nous souhaitons réaliser doit disposer d'une certaine généricité (indépendance par rapport aux applications assistées), ces différentes origines ne posent pas de problème concernant l'homogénéité de notre corpus. L'emploi d'un vocabulaire très spécifique à une application particulière permet parfois de retrouver a posteriori l'origine de certaines requêtes, mais la formulation générale des requêtes n'est elle pas affectée (par exemple, la phrase "comment faire pour arrêter le compteur ?" a une structure globale identique à "comment faire pour s'inscrire au GT ACA ?").

[3]Environ 50 sujets ayant réalisé une session représentant de 10 à 50 requêtes en environnement contrôlé sur les 3 premières applications décrites, et 30 personnes ayant participé à la campagne "Marco", ouverte sur le web à tous les usagers du GT ACA entre janvier et mars 2006.

[4]Groupe de Travail sur les Agents Conversationnels Animés - `http://www.limsi.fr/aca/`

N°	Phrases du corpus	N°	Phrases du corpus
1	**appuies** sur le bouton quitter	8	j'ai été surpris qu'il manque une fonction d'annulation
2	**clickersur** le bouton back	9	ça serait mieux si on pouvait aller directement au début
3	bon, reviens à **l apage** d'accueil	10	**auf viedersen**
4	**a** quoi sert cette fenêtre**,**	11	ca marche **:-)**
5	**c quoi** le GT ACA	12	Quel genre de musique tu aimes ?
6	le bouton "fermer" et le bouton "quitter" ont le même fonctionnement ?	13	bon à rien !
7	je ne **vosi** aucune page de **demso** ! !	14	j'aime tes cheveux Léa

TAB. 1 – Exemples de phrases du corpus DAFT avec différentes sortes d'erreurs en gras

2.3 Vue du corpus

La table 1 expose un extrait du corpus DAFT, faisant apparaître certaines de ses caractéristiques :
– beaucoup de phrases sont *bruitées* (expressions orales, fautes d'orthographe, de syntaxe et de grammaire, langage SMS…), et certaines erreurs sont non triviales à traiter avec des outils classiques de TALN.
– il ne se présente *pas* comme une succession de dialogues homme-machine, mais plutôt comme une liste de phrases employées par les usagers (questions, ordres ou remarques à l'agent…). En effet, on constate que dans le cadre de la Fonction d'Assistance, les interactions dialogiques se limitent essentiellement à un seul et unique tour de parole (commande-action, question-réponse…), et peuvent donc être traitées de manière *isolée*[5].

2.4 De la nécessité du corpus

Bien qu'il n'y ait pas de corpus d'assistance similaires aisément disponibles en français, on est en droit de s'interroger sur la nécessité de constituer un nouveau corpus : est-ce qu'un corpus dans un domaine connexe tel que le dialogue homme-machine orienté tâche n'aurait pas été suffisamment proche pour nous convenir ?

Pour répondre à cette question, et ainsi justifier la nécessité d'un corpus particulier, on peut suivre la méthode de comparaison statistique de corpus présentée dans (Ripoche, 2006) pour comparer le corpus DAFT à plusieurs corpus de dialogues homme-homme orientés tâche par l'étude de leurs profils interactionnels.
On appelle profil interactionnel d'un corpus une représentation sous forme d'histogrammes de la répartition des différents actes de dialogue (au sens de (Searle, 1969)) au sein de celui-ci (cf. fig. 2). Les trois corpus de référence choisis pour effectuer cette comparaison sont Switchboard (Jurafsky *et al.*, 1998) (200 000 énoncés de conversations téléphoniques orientées tâche annotés manuellement), MapTask (Carletta *et al.*, 1996) (128 dialogues visant à reconstruire une carte par placement de points de repère) et Bugzilla (Ripoche, 2006) (1 200 000 commentaires issus de 128 000 rapports de défauts établis lors du développement de la suite logicielle de la Fondation Mozilla).[6]

[5]Ce point essentiel de notre approche nous a été confirmé par N. Carbonell, et sera discuté en détails en 2.5.

[6]Switchboard et MapTask, de part leur nature orale, sont naturellement plus riches en nombre de mots que des corpus écrits (Kelly & Chapanis, 1977), mais la proximité des activités a primé sur cette différence de nature dans notre choix de les employer pour cette comparaison (en l'absence de corpus écrits équivalents connus).

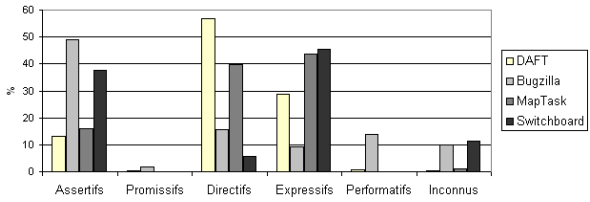


FIG. 2 – Comparaison des profils interactionnels de corpus d'assistance dialogique

Nous avons donc effectué une conversion des taxonomies d'actes de dialogue de ces différents corpus vers une taxonomie commune en 5 actes, tandis qu'un sous-ensemble au 1/5^{e} du corpus DAFT a été manuellement annoté directement dans cette même taxonomie. Certains actes de dialogues très spécifiques (par exemple les "self-talk" de Switchboard) sont difficiles à convertir ; les résultats ne sont donc pas parfaits et leur interprétation doit être considérée avec certaines précautions. Toutefois, certaines caractéristiques suffisamment nettes distinguent le corpus DAFT des trois autres :

– une présence majoritaire (57 %) de *directifs*, s'expliquant par un nombre élevé d'ordres directs ou de questions à l'agent. Bien qu'orientés tâche, les autres corpus mettent en jeu uniquement des interlocuteurs humains, et il est vraisemblable que le fait de s'adresser à la machine (même via un agent conversationnel) tend à rendre les requêtes plus directes car les usagers supposent que l'agent n'est pas capable des mêmes inférences qu'un être humain.
– un nombre assez faible d'*assertifs* (13 %), l'utilisateur exprimant bien plus son état d'esprit (29 %) par rapport à des faits que ces mêmes faits de manière neutre et "objective" comme c'est le cas par exemple dans le corpus Bugzilla ci-dessus.
– quelques *promissifs* sont présents (1 %) mais marginaux, ce qui s'explique par la nature de la relation agent-utilisateur, car si l'agent est par essence aux ordres de l'utilisateur, ce dernier ne se sent que rarement engagé envers son agent assistant (même pour suivre ses conseils).

Ces divergences entre corpus portant sur un même thème (l'assistance dialogique à une tâche) démontrent que notre hypothèse était justifiée et que notre champ d'étude se distingue suffisamment des champs connexes pour que nous ayons besoin de disposer d'un corpus propre.

2.5 L'assistance : un registre de langue ?

Le fait que la plupart des interactions ne nécessitent qu'un seul tour de parole rapproche finalement davantage notre projet des interfaces d'Interaction en Langue Naturelle (NLI en anglais, comme défini par (Androutsopoulos & Aretoulaki, 2003)) que des systèmes de dialogue à proprement parler, ce qui peut s'expliquer par le fait que le domaine de l'assistance est relativement circonscrit. Il pourrait éventuellement être considéré comme constituant un sous-langage au sens de (Kittredge, 1982) (au même titre que les bulletins météorologiques ou la biologie) ou tout du moins comme un registre particulier de langue (Biber, 1995), dans la mesure où si l'on fait abstraction du clavardage entre l'utilisateur et l'agent qui n'est pas notre centre d'étude principal, on constate globalement que :

– le vocabulaire employé est assez pauvre et fortement lié à l'application assistée (cf. phrases 1-9 de la table 1).
– les classes lexicales les plus fréquentes sont peu nombreuses car essentiellement en rapport avec des éléments de l'interface graphique (boutons, champs texte…) et avec des actions

standards (ajouter, modifier, déplacer…).
– les structures de phrases, particulièrement dans les sous-corpus d'assistance, sont assez prototypiques et très différentes de ce que l'on trouve dans la langue naturelle générale[7].

3 Catégorisation et caractérisation du corpus DAFT

3.1 Catégorisation des activités conversationnelles

Lors de la phase de recueil du corpus, les sujets humains ont été informés qu'ils devaient réaliser certaines tâches, en faisant appel si nécessaire à un agent (non humain) présent dans l'application pour les assister. Les sujets pouvaient néanmoins agir et s'exprimer de manière non contrainte, et divers comportements ont pu être observés, l'utilisateur se détournant parfois complètement de sa tâche initiale. Finalement, il apparaît donc que de nombreuses phrases recueillies ne relèvent pas vraiment du domaine de l'assistance (cf. table 1). Nous nous sommes par conséquent intéressés à identifier les différentes activités conversationnelles réellement présentes dans le corpus.

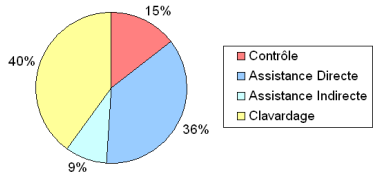


FIG. 3 – Répartition du corpus en sous-corpus par activités

Pour cela, nous avons extrait aléatoirement des phrases du corpus de manière à former deux sous-ensembles de taille égale au dixième de la taille totale du corpus. Dans le premier sous-ensemble, les phrases ont été regroupées manuellement par activités similaires, en s'intéressant notamment au thème des requêtes (application, agent, utilisateur…) et à leur nature (ordre, question, compliment…). On a finalement obtenu ainsi quatre grandes catégories de tailles inégales (cf. fig. 3). Ensuite, connaissant ces quatre activités principales, on a traité le deuxième sous-ensemble en classifiant manuellement les phrases de celui-ci dans une des activités définies précédemment. La répartition ainsi déterminée était très proche de celle trouvée sur le premier sous-ensemble, ce qui nous laisse penser qu'on peut raisonnablement généraliser ce résultat à l'ensemble du corpus (la figure 3 présente les résultats obtenus en faisant la moyenne des deux répartitions trouvées).
On peut par conséquent considérer que notre corpus est divisible en quatre "sous-corpus", correspondant chacun à des types d'activités distinctes (les phrases données en exemples sont celles de la table 1) :

1. **activité de contrôle :** corpus constitué de *commandes*, afin que l'agent agisse lui-même sur l'application *(phrases 1-3)*.
2. **activité d'assistance directe :** regroupant des *demandes d'aide* explicitement formulées comme telles par l'utilisateur *(phrases 4-6)*.

[7]Comme l'a montré une autre comparaison effectuée par rapport à un extrait du corpus généraliste Multitag, fourni par P. Paroubek.

Sous-corpus	Contrôle	Assist. directe	Assist. indirecte	Clavardage
Moyenne	5,44	8,01	9,90	6,01
Écart-type	3,36	3,54	3,30	3,62

TAB. 2 – Répartition des phrases par longueur (mots) dans les sous-corpus

3. **activité d'assistance indirecte :** corpus formé d'*opinions* sur l'application qui constituent des demandes d'aide sous-entendues, probablement perceptibles uniquement au niveau pragmatique *(phrases 7-9)*.
4. **activité de clavardage :** réunissant le reste des interactions essentiellement centrées sur l'agent ainsi que des expressions métalinguistiques, phatiques[8] et de backchanneling[9] *(phrases 10-14)*.

L'existence des sous-corpus de contrôle et de clavardage démontre que l'utilisateur attend non seulement d'un ACA qu'il l'aide à utiliser une application, mais aussi qu'il soit capable d'agir lui-même sur celle-ci, ainsi que de répondre à des commentaires annexes indépendants de la tâche à accomplir où l'agent devient lui-même le centre d'intérêt de l'utilisateur (ce phénomène s'expliquant essentiellement par la présence d'une représentation visuelle de l'agent).

3.2 Méthodes de caractérisation des sous-corpus

Les quatre sous-corpus ont été catégorisés précédemment uniquement par une annotation *manuelle*, mais il serait souhaitable de pouvoir automatiser cette classification afin d'analyser spécifiquement les activités propres à chaque sous-corpus. On envisage alors trois méthodes de caractérisation possibles de ceux-ci :
– une étude de la distribution de la longueur des phrases des sous-corpus.
– une étude des profils interactionnels des sous-corpus, tels que définis en 2.4.
– une étude de la sémantique des phrases par analyse de leur retranscription sous forme de requêtes formelles.

3.2.1 Caractérisation par la longueur des phrases

On observe une certaine disparité de longueur des requêtes, les requêtes de contrôle semblant globalement assez courtes comparées aux requêtes d'assistance (cf. table 2), et on peut approximer les répartitions des sous-corpus de contrôle et d'assistance indirecte par une loi normale (test de χ^2 avec un seuil de tolérance de 1%). Néanmoins, les écart-types trop importants ($\sigma \approx 3,5$) disqualifient en pratique cette méthode de classification.

3.2.2 Caractérisation par l'analyse des profils interactionnels

On distingue sur la figure 4 certaines différences de profils interactionnels assez nettes entre les sous-corpus et par rapport au profil générique du corpus DAFT (rappelé en gris foncé), notamment pour distinguer l'assistance directe (avec une forte majorité de directifs et quelques expressifs) de l'assistance indirecte (une majorité d'assertifs et des expressifs). En revanche, les

[8]Pour maintenir le contact communicatif avec l'agent : « pas vrai lea ? », « tu dors ou quoi ? »…
[9]Pour marquer son accord aux propos du locuteur et l'inciter à continuer : « Bon. », « ok ok »…

profils interactionnels des sous-corpus de contrôle et d'assistance directe sont assez similaires. Cette méthode présente donc un certain intérêt mais ne peut être utilisée de manière unique. En outre, automatiser la détection de ces actes de dialogue n'est pas non plus trivial.

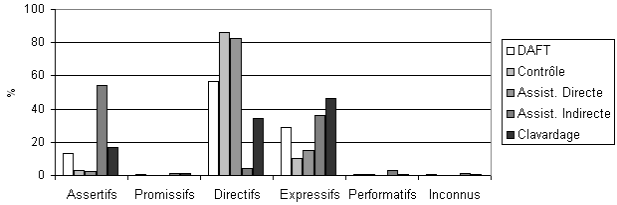


FIG. 4 – Comparaison des profils interactionnels des sous-corpus du corpus DAFT

3.2.3 Caractérisation par analyse des requêtes formalisées

Notre principale motivation pour constituer ce corpus était de pouvoir s'en servir comme base pour la définition d'un langage formel adapté pour exprimer la sémantique des requêtes langagières. La syntaxe de ce langage, présentée en détails dans (Bouchet, 2006), distingue des modalités (la possibilité, l'obligation, le savoir...), des prédicats d'actions (modifier, déplacer, actionner...) et des références (le tableau, le petit bouton rouge...) qui peuvent s'imbriquer entre elles sous la forme :

$$\mathbf{M_1(\ldots M_n(c_1 = P_1(c'_1 = R_1, \ldots, c'_l = R_l), \ldots, c_m = P_m(\ldots)) \ldots)}$$

Expression dans laquelle : M_1 - M_n sont des **modalités**,
P_1 - P_m sont des **prédicats**,
c_1 - c_m & c'_1 - c'_l sont des intitulés de champs typés (par exemple : objet, personne, lieu...),
R_1 - R_l sont des **références** issues de la requête langagière.

Exemple : La phrase 6 de la table 1 pourra ainsi s'écrire sous la forme :
CHECK(NEG(DIFFERENCE(between="le bouton 'fermer'", and="le bouton 'quitter'", is="fonctionnement")))

Si l'on s'intéresse alors au nombre de modalités ou de prédicats présents dans chacun des sous-corpus, on constate des différences assez nettes (cf. figure 5) :
– le contrôle possède un nombre élevé de prédicats par phrase (0.97) et peu de modalités (0.22).
– l'assistance directe contient beaucoup de modalités (2.41) ainsi qu'un nombre moyen de prédicats (0.54).
– l'assistance indirecte diffère peu de l'assistance directe avec 0.59 prédicats par phrase et légèrement moins de modalités (2.22), avec toutefois une part plus forte de phrases à modalité unique (18% contre 5%).
– le clavardage (limité aux phatiques pour cette étude) est extrêmement pauvre en prédicats (0.08) et assez peu de modalités (0.86).

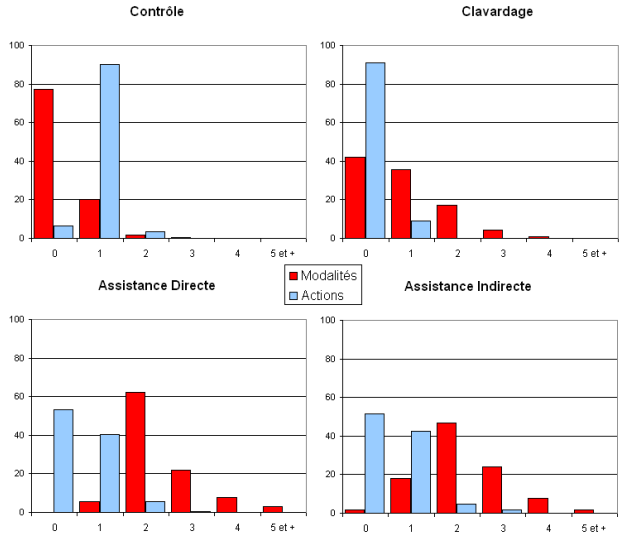


FIG. 5 – Nombre de modalités et prédicats par phrase dans chacun des sous-corpus

3.2.4 Conclusion de l'étude

Parmi les trois méthodes testées, aucune ne permet d'identifier parfaitement l'activité correspondant à une requête. On constate en effet que :
– la longueur des phrases n'apporte pas d'informations discriminantes,
– la comparaison des profils interactionnels permet une discrimination efficace des deux formes d'assistance (directe et indirecte) qui nous intéressent dans le cadre du projet DAFT,
– la comparaison des requêtes mises sous forme formelle permet de distinguer le contrôle, l'assistance et le clavardage.

Pour classifier automatiquement les requêtes dans un des sous-corpus, on pourra donc envisager de combiner ces deux dernières méthodes.

4 Conclusion et perspectives

La Fonction d'Assistance, dans le cadre des CHS, constitue un registre de langue particulier, et se distingue d'activités connexes comme le dialogue homme-homme orienté tâche. Nous avons constitué un corpus composé de requêtes d'assistance recueillies dans le cadre de diverses applications et d'autres requêtes construites en situation, ce qui nous a permis d'avoir un ensemble globalement représentatif de l'assistance. Nous avons montré que ce corpus recouvre en réalité quatre activités distinctes, identifiables par une combinaison de méthodes statistiques classiques et d'une analyse nécessitant la transcription des requêtes dans un langage formel.

Depuis la réalisation de cette étude, le corpus DAFT a été complété, notamment afin de renforcer la proportion de requêtes d'assistance (qui ne représentait ici que 45% du corpus total) en incluant des requêtes relevées face à une application de nature plus complexe (Word). Il compte

ainsi désormais un peu plus de 11 000 phrases, qui nous servent actuellement de base solide pour la construction de la chaîne de traitement des requêtes langagières (analyses grammaticale et sémantique). Un large extrait sous forme brute ainsi que des phrases analysées par notre système sont disponibles en libre accès[10].

Références

ANDROUTSOPOULOS I. & ARETOULAKI M. (2003). *The Oxford Handbook of Computational Linguistics*, chapter Natural Language Interfaces, p. 629–649. Oxford University Press.

ATKINS B. & LEWIS H. (1996). *The Collins-Robert French-English Dictionary*, chapter Language in Use. Harper Collins Publishers, 1st edition.

BIBER D. (1995). *Dimensions of register variation : a cross-linguistic comparison*. Cambridge University Press.

BOUCHET F. (2006). Conception d'un langage de requêtes pour un agent conversationnel assistant. Master's thesis, Univ. Paris XI.

BUISINE S. & MARTIN J.-C. (2005). Children's and adults' multimodal interaction with 2d conversational agents. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 1240–1243, Portland, Oregon, USA : ACM Press.

CAPOBIANCO A. & CARBONELL N. (2002). Conception d'aides en ligne pour le grand public : défis et propositions. In *Actes de ERGO-IA'2002*, p. 309–335.

CARLETTA J., ISARD A., ISARD S., KOWTKO J., DOHERTY-SNEDDON G. & ANDERSON A. (1996). *HCRC dialogue structure coding manual*. Rapport interne, Univ. of Edinburgh.

JANSEN B. J. (2005). Seeking and implementing automated assistance during the search process. *Information Processing and Management*, **41**(4), 909–928.

JURAFSKY D., BATES R., COCCARO N., MARTIN R., METEER M., RIES K., SHRIBERG E., STOLCKE A., TAYLOR P. & VAN ESS-DYKEMA C. (1998). *Switchboard Discourse Language Modeling Project Final Report*. Rapport interne, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, USA.

KELLY M. J. & CHAPANIS A. (1977). Limited vocabulary natural language dialogue. *International Journal of Man-Machine Studies*, **9**(4), 479–501.

KITTREDGE R. (1982). Variation and homogeneity of sublanguages. In *Sublanguage : Studies of Language in Restricted Semantic Domains*, p. 107–137. De Gruyter.

LE GUERN K. (2004). Définition d'une architecture de médiateur pour des agents conversationnels animés. Master's thesis, Univ. Paris XI.

LESTER J. C., CONVERSE S. A., KAHLER S. H., BARLOW S. T., STONE B. A. & BHOGAL R. S. (1997). The Persona Effect : Affective impact of animated pedagogical agents. In *Proc. of CHI'97*, p. 359–366, New York, NY, USA : ACM Press.

MOLINSKY S. J. & BLISS B. (1994). *Inventory of functions and conversation strategies*, In *Communicator : The Comprehensive course in functional English*, p. 177–187. Prentice Hall.

RIPOCHE G. (2006). *Sur les traces de Bugzilla*. PhD thesis, Univ. Paris XI.

SADEK D., BRETIER P. & PANAGET E. (1997). Artimis : Natural dialogue meets rational agency. In *IJCAI (2)*, p. 1030–1035.

SANSONNET J.-P., LE GUERN K. & MARTIN J.-C. (2005). Une architecture médiateur pour des agents conversationnels animés. In *Actes de WACA'01*, p. 31–39.

SEARLE J. R. (1969). *Speech Acts : An essay in the Philosophy of language*. Cambridge, new edition.

[10] `http://www.limsi.fr/~jps/research/daft/index.html`

Extraction endogène d'une structure de document pour un alignement multilingue

Romain BRIXTEL
Laboratoire GREYC, Université de Caen-Basse-Normandie
Campus II, 14032 Caen Cedex
rbrixtel@info.unicaen.fr

Résumé. Pour des raisons variées, diverses communautés se sont intéressées aux corpus multilingues. Parmi ces corpus, les textes parallèles sont utilisés aussi bien en terminologie, lexicographie ou comme source d'informations pour les systèmes de traduction par l'exemple. L'Union Européenne, qui a entraîné la production de document législatif dans vingtaine de langues, est une des sources de ces textes parallèles. Aussi, avec le Web comme vecteur principal de diffusion de ces textes parallèles, cet objet d'étude est passé à un nouveau statut : celui de document. Cet article décrit un système d'alignement prenant en compte un grand nombre de langues simultanément (> 2) et les caractéristiques structurelles des documents analysés.

Abstract. For many reasons, the multilingual corporas have interested various communities. Among these corporas, the parallel texts are used as well in terminology, lexicography or as a source of informations for example-based translations. The European Union, which involved the production of legislative documents, generates these parrallel texts in more than twenty languages. Also, with the Web as a vector of diffusion, we can wonder if these parallel texts can be treated as documents. This article describes a alignment system taking account a great number of languages (> 2) and the structural characteristics of the analyzed documents.

Mots-clés : alignement multilingue, corpus parrallèles, multitextes, multidocuments, extraction de structures, alignement endogène.

Keywords: multilingual alignment, parallel corpora, multitexts, multidocuments, extraction of structures, endogenous alignment.

1 Introduction

L'alignement est une opération qui consiste à relier des unités qui se correspondent dans des textes parallèles. Des éléments peuvent alors être appariés suivant le grain d'analyse (paragraphes, phrases ou d'autres unités plus fines telles que les mots) sur lequel le système d'alignement s'appuie.

L'intérêt croissant porté à cet axe de recherche est lié à l'augmentation de la production des corpus multilingues regroupant des textes et leurs traductions (textes parallèles ou multitextes), ainsi que leur accessibilité de plus en plus aisée.

Dans cet article, nous présentons tout d'abord un état de l'art décrivant des méthodes d'aligne-

ments à différents niveaux de granularité. Ensuite, nous aborderons la méthode d'alignement multilingue (près d'une vingtaine de langue) implémentée en détaillant les hypothèses utilisées et en présentant les différents alignements générés

2 État de l'art

Deux grains principaux nous intéresse : l'alignement de phrases ainsi que l'alignement d'éléments sous-phrastiques tel que les mots, termes ou expressions.

2.1 Alignement de phrases

Pour l'alignement au grain phrase, deux méthodes principalement utilisées sont à retenir.

(Kay, 1988) part de l'hypothèse suivante : deux phrases sont alignables si les mots qui les composent sont en correspondance. En supposant que certaines phrases sont alignées au début du traitement (habituellement les premières et dernières phrases des documents du multidocument), le processus propose des candidats de phrases à aligner. Aussi, l'algorithme considère leurs places dans le document. On exclut la possibilité d'aligner une phrase se situant au début d'un document avec une autre à une place très eloignée dans l'autre document. Le processus d'alignement est itératif. L'algorithme valide les alignements phrastiques en satisfaisant au maximum les équivalences d'ensembles de mots faites dès le début et en cherchant des phrases équivalentes avec un rang très proche. Chaque nouvelle paire de phrases alignées restreint les domaines d'équivalences entre les mots de langues différentes.

(Gale & Church, 1993) se basent sur la propriété suivante : des phrases sont alignées si leurs longueurs relatives sont équivalentes. La méthode part du principe qu'une phrase longue écrite dans une langue sera traduite par une phrase longue dans une autre langue. La longueur des phrases est alors relative à la longueur moyenne, en nombre de caractères, des phrases de la langue dans laquelle elles sont écrites. De la même façon que la méthode de (Kay, 1988), on admet ici que les phrases alignées ont un rang très proche dans les deux documents analysés. Sont alors envisagés les alignements $1:2$ (une phrase correspond à deux autres), $2:1$, $1:0$ et $0:1$. Ainsi, en l'absence de candidat à l'alignement pour une phrase donnée, la méthode propose de regarder si elle correspond à plusieurs phrases suivant le même critère de longueur, en se limitant aux possibilités d'alignement décrits précédemment. Les possibilités de fusions ou d' omissions de phrases sont inclues par raison pratique en terme de simplification de calcul ; tout en suivant une certaine logique du processus de traduction.

Des méthodes d'ancrages lexicaux préalables sont utilisées en complément pour affiner les résultats. Ainsi, (Debili & Sammouda, 1992), (Simard *et al.*, 1992), (Melamed, 1999) utilisent les similarités graphiques entre les langues (cognates) pour mettre des mots en correspondance. En suivant l'hypothèse de (Kay, 1988), cette technique améliore les alignements phrastiques

Les hypothèses sous-jacentes aux méthodes d'alignement de type (Kay, 1988) et (Gale & Church, 1993), pour un alignement d'une paire de texte au grain phrase, peuvent être résumées de la façon suivante (Véronis, 2000) :

– l'ordre des phrases dans les deux textes est identique ou très proche ;
– les textes contiennent peu de suppressions ou d'adjonctions ;

– les alignements $1 : 1$ sont très largement majoritaires et les rares alignements $m : n$ sont limités a de petites valeurs de m et n (≈ 2).

Même si ces hypothèses paraissent légitimes, les systèmes d'alignement se basant dessus rencontrent de grandes difficultés si les documents comportent des différences structurelles (par exemple : date figurant au début d'un texte et à la fin d'un autre).

2.2 Alignement sous-phrastique

Dans le cadre de l'alignement en phrases, la mise en correspondance de mots n'est pas le but premier mais seulement un amorçage : on se satisfait ici d'un alignement grossier de mots. Si l'ordre des phrases de deux traductions est plus ou moins respecté, il en est autrement dès que l'on se situe au niveau de la phrase. L'insertion d'adverbes ou d'adjectifs dans une expression ou encore la permutation entre deux propositions d'une phrase ne permet pas de faire les mêmes suppositions que pour l'alignement au niveau de la phrase. De plus, les phénomènes d'agglutination et de flexions ébranlent les méthodes statistiques basées sur la comparaison de chaînes de caractères constantes

L'alignement en éléments sous-phrastiques consiste à détecter ces éléments puis à les mettre en correspondance. Or, séparer ces deux étapes est délicat car déterminer la nature exacte des éléments sous-phrastiques en jeu (par exemple : mots, expressions, chunk, propositions) dépend aussi bien de la langue cible que de la langue source (par exemple : "rideau de fer" se traduit par "Eisenvorhang" en allemand et "iron curtain" en anglais). L'utilisation d'outils statistiques est donc extrêmement délicat dans de telles conditions et choisir un de ces outils se révèle être du "cas par cas" en fonction des langues à aligner. De plus, la plupart des expressions sont seulement "semi-figées" et peuvent être altérées par certaines opérations (flexion, insertion d'adjectifs, passivation...).

(Giguet, 2005) considère le problème d'alignement sous-phrastique comme étant celui de l'appariement de suites de mots ayant une répartition similaire à travers les textes analysés. Ce choix de traitement statistique influe sur les résultats. La répartition de mots ayant une graphie identique à travers le document dépend des caractères flexionnel et agglutinant de la langue considérée. Une analyse morpho-endogéne des documents permet alors des alignements satisfaisants entre des documents anglais et grecque ; mais ne résout pas la comparaison entre deux documents de langues ayant un caractère agglutinant différent, tel que l'anglais et le finnois.

(Zimina, 2004) applique une approche textométrique sur des corpus multilingues. Sur un texte parallèle français-anglais-russe, elle montre que lorsque qu'un mot est doté d'un large éventail de sens dans le corpus, la comparaison des fréquences totales des formes graphiques ne constitue pas toujours une bonne indication pour l'appariement (par exemple : la correspondance d'un mot polysémique comme "droit"). L'étude des formes graphiques en contexte (avec leur voisinage, les mots les entourant) permet alors de lever certaines des ambiguïtés résiduelles. Au contact d'autres mots associés sur l'axe syntagmatique, différents composants du sens du mot sont activés et il devient possible d'en tenir compte lors de l'appariement. La répartition d'un mot et la répartition des mots voisins sont utilisés comme éléments de comparaison.

La difficulté principale de l'alignement à ce grain réside dans la difficulté à cerner les éléments que l'on veut aligner, les phénomènes linguistiques sont plus nombreux (sans pour autant que l'un se détache des autres) et sont différents suivant les langues. L'introduction de connaissances linguistiques est alors relativement coûteuse et l'on devient dépendant des langues traitées.

3 Quel grain pour un alignement multilingue ?

Un des points qui émerge de cette vision globale de l'alignement est que les traitements à un grain affectent ceux à d'autres grains. D'une part (Kay, 1988) se base sur ce que l'on peut appeler un alignement grossier au grain mot pour un alignement de phrases, d'autre part un alignement à un grain phrastique peut être un préalable à l'alignement de mots ou de suite de mots ((Zimina, 2004), (Giguet, 2005)).

Les indices majoritairement utilisés [1] proviennent d'une vision de l'alignement comparable à celle abordée sur les problèmes de traitements de flux de caractères ou de séquence de mots. D'autre part, l'alignement a été abordé comme un problème de découpe de textes et de mise en correspondance des morceaux découpés (typiquement, découper le texte en phrases puis trouver les correspondances entre elles, ou trouver la couverture maximale entre les éléments issus de la découpe de deux phrases alignées). La hiérarchie $texte \rightarrow paragraphe \rightarrow phrase \rightarrow expression \rightarrow mot$ est alors la plus utilisée pour définir les grains qui seront traités. L'originalité de la méthode proposée dans cette article est de se placer en parallèle de cette hiérarchie sans pour autant l'ignorer.

Le Web est le plus grand vecteur de diffusion de ces textes, il les affecte alors naturellement. Voir l'alignement autrement qu'en envisageant exclusivement les objets analysés comme des flux de caractères va dans ce sens : l'utilisation de liens hypertextes, d'images, de tableaux ou d'autres applications de mise en forme matérielle (MFM) telles que des marques de graisse et d'emphase, affecte le statut des traductions disponibles. En considérant ces traductions comme des documents, les multitextes décrits dans la littérature peuvent être abordés comme des "multidocuments".

Très peu de techniques prennent en compte l'aspect réellement multilingue des multidocuments (qui se résument dans la littérature majoritairement à des multidocuments de deux documents, ou bi-documents) qui est l'essence même des multidocuments. On peut se demander si à trop s'attacher à peu de langues, nous n'obtenons pas des méthodes de traitement ad hoc : le multilinguisme est une force qui permet l'abstraction alors que se cantonner à un nombre de langues réduit oriente les méthodes de traitement que l'on peut appliquer. (Simard, 2000) montre que l'ajout de langues rends plus fiable les alignements générés automatiquement, mais sa méthode se résume plus à l'utilisation conjointe de plusieurs bi-documents qu'un multidocument dans son ensemble.

Le nombre de langues peut agir comme un filtre suffisamment fort pour contraindre l'apparition de résultats. Aussi, même si l'on veut tendre vers un idéal multilingue maximal, il est toujours possible d'exclure une langue d'un multidocument du traitement. Le fait de remarquer qu'une langue ou qu'un type de document est "récalcitrant" à une hypothèse que l'on a formulé peut nous apprendre beaucoup sur le document/la langue (ou le groupe de langues/documents) exclu et la façon d'aborder le multidocument.

Collectant et diffusant des communiqués de presse de l'Union Européenne sous format électronique à travers leur site Web, Europa [2] permet de récupérer des multidocuments comprenant des documents écrits dans plus de vingt langues différentes. A partir d'un corpus d'étude [3] extrait de

[1] Longueur des phrases, contenu des phrases (principalement : mots ou suite de mots), position des phrases ou d'éléments sous-phrastiques (mots ou suite de mots), distance graphique d'éléments sous-phrastiques (cognates), patron linguistique d'extraction propre à une langue ou à une famille de langue, etc.

[2] `http://europa.eu/`

[3] 63 multidocuments de plus de 16 documents/langues chacun.

ce site, nous nous orientons vers un alignement prenant en compte simultanément le maximum de langues en parallèle à la hiérarchie $texte \rightarrow paragraphe \rightarrow phrase \rightarrow expression \rightarrow mot$. Il faut alors trouver les invariants qui peuvent être détectés dans tous les documents pour définir les bases d'un treillis d'informations robuste.

4 La MFM : un invariant de structure multilingue ?

Les documents extraits d'Europa se présentent sous la forme de documents XHTML [4]. De nombreuses traces non-textuelles peuvent y être repérées via la MFM telles que les tableaux, les séparations horizontales (sauts de ligne via la balise
 ou les traits horizontaux via la balise <hr/>), l'application de graisse et d'italique, voir des liens hypertextes [5]. L'avantage de ces marques est qu'elles mettent en valeur des zones à l'intérieur des documents. Une phrase ayant une MFM particulière, en plus d'être mise en valeur dans son intégralité, met aussi en avant chaque élément dont elle est composée (chunks, propositions, mots ou caractères). Une approche consiste à se demander si ces marques peuvent scinder le document en plusieurs parties. Par exemple, en détectant une ligne de séparation dans un document il est possible de considérer deux zones, celle avant la ligne et celle située après. Si cette séparation est présente dans tous les documents du multidocument, alors nous pouvons mettre en relation les zones ainsi dégagées.

Il parait douteux d'effectuer une dichotomie sur un document suivant l'apparition d'un élément sous-phrastique mis en valeur (par exemple : un mot mis en gras). Au grain sous-phrastique, sa position dans la phrase est affectée par les règles de construction propre à la langue du document. A ce grain, ces indices sont des marques caractérisant les grains dans lesquels ils sont inclus. Comme les cognates, ils sont détectés pour marquer ces grains (les phrases pour les cognates) et les aligner ; ils ne sont pas utilisés pour diviser le document.

Si le grain phrase n'est pas approprié, nous considérons un grain supérieur. Un alinéa [6] peut ainsi être mis en gras afin qu'il soit assimilé à un titre de partie. La mise en évidence de passage peut être une finalité en soit ou un moyen d'organiser le discours. Par l'utilisation de titres, de résumés ou de passages ayant une MFM différente de celle appliquée au corps de texte, le document est découpé en différents segments.

L'alignement peut être vu comme l'extraction d'équivalences sémantiques. Considérer cet usage de la MFM comme un vecteur de sens préservé dans le processus de traduction nous amène à exploiter ces marques en tant qu'invariant entre les documents de différentes langues.

Cette segmentation à un grain plus élevé que la phrase permet de restreindre les espaces de recherche [7] dans le cadre de la recherche d'équivalences sémantiques entre les documents d'un multidocument. Par exemple, l'alignement phrastique suppose que chaque phrase est alignable avec une autre ayant une position très proche dans le texte. Dégager des zones permettrait de cibler nos recherches pour dégager les équivalences : une phrase appartenant à une zone est potentiellement alignable avec les phrases des autres documents appartenant à la même zone du texte. Et si cette méthode parait viable dans le cadre de l'alignement phrastique, elle peut

[4] eXtensible HyperText Markup Language, recommandation w3c `http://www.w3.org/MarkUp/#recommendations`

[5] Les liens hypertextes ne sont pas des marques de MFM mais leurs mises en relief en font parti.

[6] `http://atilf.atilf.fr/` — Texte compris entre deux retours à la ligne.

[7] « Diviser pour régner ».

aussi l'être dans les grains contenus dans les zones de recherches détectées (les grains contenus peuvent être des paragraphes, des mots, des expressions...).

4.1 Caractérisation des alinéas par la MFM

Nous partons d'un découpage en alinéas des documents. Sur des documents XHTML, ces alinéas peuvent être détectés grâce aux balises de bloc (telles que <p>, <h1>, <h2>,) en opposition aux balises en ligne (comme , ,). D'autres méthodes comme celles utilisant les techniques de reconnaissances de texte [8] peuvent être utilisées pour arriver à cet objectif. Seules les MFM appliquées sur un alinéa dans son ensemble nous serviront à diviser les documents en parties.

La première étape consiste à détecter la MFM utilisée pour chaque alinéa afin de permettre une segmentation du document. Or, une MFM n'est pas systématiquement appliquée tout le long d'un alinéa, comme nous pouvons le remarquer sur l'exemple ci-dessous [9] :

[…] Ir iekl,autas TRIPS *un* PVO *prasi-bas, […]*

Les acronymes "TRIPS" et "PVO" ont été mis en valeur dans cet alinéa en modifiant la MFM qui a été attribué en majorité. Si nous considérons les MFM appliquées à un alinéa seulement si elles sont présentes à tout endroit du texte, nous ne caractérisons pas correctement cet alinéa. Pour pallier ces problèmes, nous considérons la MFM d'un alinéa de la façon suivante :

– soit mfm_{deb} l'ensemble des MFM appliquées au premier mot de l'alinéa ;
– soit mfm_{fin} l'ensemble des MFM appliquée au dernier mot de l'alinéa ;
– la MFM caractérisant l'alinéa est définie par $\text{mfm}_{al} = \text{mfm}_{deb} \cup \text{mfm}_{fin}$.

Nous évitons les problèmes cités ci-dessus ainsi que ceux pouvant être provoqués par l'usage de lettrine ou encore ceux provoqués par un usage important d'une MFM dans un alinéa, comme par exemple dans le cas d'une citation en italique et occupant une grande partie de l'alinéa dans lequel elle est située.

4.2 Découpage du document par la MFM

La figure (FIG. 1) présente une visualisation des documents estonien (et), italien (it) et néerlandais (nl) [10] segmentés en alinéas . La segmentation est faite sur 20 langues même si nous n'en montrons ici que 3. Un alinéa est représenté par l'identifiant de sa MFM.

Chaque numérotation des identifiants est interne à chaque langue. Par exemple, un document peut être écrit en majorité en italique là où un autre est ecrit en romain : l'invariant considéré n'est donc pas un invariant de forme mais de structure.

De cette visualisation, nous pouvons pour un document donné dégager deux types d'alinéas : les alinéas nombreux et contigus (contenant le corps du texte) et les autres (révélateurs d'une

[8] OCR

[9] Extrait du document letton IP\05\1659 `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/1659&format=HTML&aged=1&language=LV&guiLanguage=en`

[10] Extrait du multidocument IP\05\817. `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/817&format=HTML&aged=1&language=IT&guiLanguage=en`

id.	extrait de l'alinéa avec sa MFM
a	Bruxelles 30 giugno 2005
b	**2006 - […] professionale**
c	***La Commissione […] lavoro.***
a	Lavorare in un […] di 10 anni.
a	Vladimír Spidla […] lavorare".
a	Su un bilancio […] professionale.
a	Nel 2006 […] commissari.

alignement au grain document		
et		1 1 2 3 1 1 1
it		a b c a a a a
nl		$\alpha\ \alpha\ \beta\ \gamma\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha$

FIG. 1 – Alinéas du document it et identifiants de MFM d'alinéa pour les documents et, it et nl

structure). En faisant cette distinction, nous regroupons automatiquement les suites contigues d'alinéas de même MFM. Les alinéas mis en valeur sont alors ceux qui brisent ces suites, que nous alignons aussi entre eux (alinéas "b" et "c" pour le document it - FIG. 2).

id.	extrait de l'alinéa avec sa MFM
a	Bruxelles 30 giugno 2005
b	**2006 - […] professionale**
c	***La Commissione […] lavoro.***
a	Lavorare in un […] di 10 anni.
a	Vladimír Spidla […] lavorare".
a	Su un bilancio […] professionale.
a	Nel 2006 […] commissari.

alignement 1										
et		1 1		2		3		1 1 1		
it		a		b		c		a a a a		
nl		$\alpha\ \alpha$		β		γ		$\alpha\ \alpha\ \alpha\ \alpha\ \alpha$		

FIG. 2 – Découpage du document italien pour l'alignement 1

Cependant, considérer seulement les suites d'alinéas comme précédemment (FIG. 2) ne révèle pas toutes les équivalences. Si nous alignons les titres entre eux et les sous-parties entre elles de façon indépendante, nous obtenons un découpage qui restreint simplement les espaces de recherche. Nous pouvons proposer manuellement les équivalences suivantes (FIG. 3) qui ne sont pas dévoilées par l'alignement automatique opéré précédemment.

id.	extrait de l'alinéa avec sa MFM
a	Bruxelles 30 giugno 2005
b	**2006 - […] professionale**
c	***La Commissione […] lavoro.***
a	Lavorare in un […] di 10 anni.
a	Vladimír Spidla […] lavorare".
a	Su un bilancio […] professionale.
a	Nel 2006 […] commissari.

alignement 2									
et		1 1		2		3 1 1 1			
it		a		b		c a a a a			
nl		$\alpha\ \alpha$		β		$\gamma\ \alpha\ \alpha\ \alpha\ \alpha\ \alpha$			

FIG. 3 – Découpage du document italien pour l'alignement 2

2, b et β représentent des titres dans les trois documents. L'équivalence montrée en (FIG. 3) dévoile le fait que les parties qui suivent chaque titre peuvent être aussi alignées. Afin de multiplier les alignements possibles sans pour autant perdre l'alignement simple vu pour l'alignement 1

(FIG. 2), nous nous orientons vers une méthode d'extraction de structure des documents du multidocument.

4.3 Structuration du document par la MFM

Le but ici n'est pas de tenter la détection de la structure logique des documents, mais de réussir à calculer une représentation de structure. Nous ne cherchons pas à savoir si un élément détecté est une signature ou un résumé même si cette étape peut servir de base à d'éventuelles applications allant dans ce sens. Nous cherchons simplement à rendre les documents comparables entre eux.

Dans l'exemple précédant 1, a et α représentent la MFM du corps de texte, les autres alinéas conditionnent alors la structure. Pour différencier ces alinéas, nous utilisons l'ordre d'apparition de ceux-ci dans le document. Dans le but de représenter la structure sous forme d'arbre, nous considérons que descendre dans la hiérarchie arborescente du document s'effectue suivant l'ordre de lecture du document en fonction des alinéas ayant une MFM différente de celle représentant le corps de texte. Aussi, deux alinéas ne représentant pas le corps de texte et ayant la même MFM sont à la même profondeur de l'arbre, au même niveau de hiérarchie. Suivant ces remarques, nous établissons que :

– Chaque alinéa ayant une MFM représentant le corps de texte
  - possède un niveau de hiérarchie dépendant des alinéas précédemment rencontrés ;
  - n'augmente pas le niveau de hiérarchie des prochains alinéas.
– Chaque alinéa ayant une MFM différente de celle représentant le corps de texte
  - possède un niveau de hiérarchie dépendant des alinéas précédemment rencontrés. Ce niveau est alors fixé pour tous les alinéas ayant la même MFM ;
  - augmente le niveau de hiérarchie des prochains alinéas.

Le tableau (FIG. 4) montre l'application de la méthode sur le document italien.

modification de niveau de hierarchie		**+1**	**+1**				
niveau de hierarchie	1	1	2	3	3	3	3
alinéas	a	b	c	a	a	a	a

FIG. 4 – Calcul des niveaux de hierarchie pour le document italien

Afin d'illustrer la mise en correspondance des niveaux de hiérarchie de deux alinéas ayant deux MFM identiques, le tableau (FIG. 5) présente l'application de la méthode de hiérarchisation sur un document ayant une structure plus complexe [11].

modification de niveau de hierarchie		**+1**	**+1**				**+1**		**+1**			
niveau de hierarchie (! :fixe les niveaux *, * :niveau fixé)	1	1	2	3	3	3	3**!**	4	3*	4	4	4
alinéas	A	B	C	A	A	A	D	A	D	A	A	A

FIG. 5 – Calcul des niveaux de hierarchie pour le document français (IP\05\606)

[11] Document français du multidocument IP\05\606. `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/606&format=HTML&aged=1&language=FR&guiLanguage=en`

Il est possible alors de générer les arborescences à partir de la hiérarchisation qui a été calculée. La racine (niveau 0) représente pour chaque arbre le grain englobant les alinéas : le document.

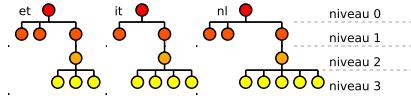


FIG. 6 – Structures générées - documents estonien, italien et néerlandais

Pour comparer les arbres entre eux, nous reprenons les critères utilisés pour l'alignement 2 (FIG. 3). Les suites connexes de feuilles sont regroupées pour permettre une comparaison entre les documents de structures équivalentes.

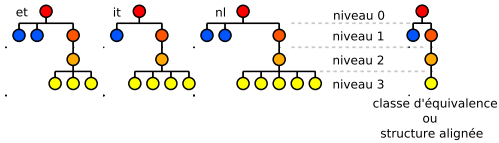


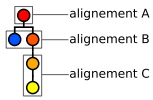
FIG. 7 – Alignement des structures

En regroupant ainsi les structures, nous obtenons une classe d'équivalence de structures (ou structure alignée). Chaque noeud et chaque feuille contiennent des suites d'alinéas alignées. Les documents sont aussi regroupés dans le même noeud (FIG. 8 alignement A). Ceci va avec l'idée que créer un multidocument équivaut à aligner des documents entre eux (ou à aligner au grain document). D'autres alignements peuvent être révélés dans les relations frère-frère (FIG. 8 alignement B) ou père-fils (FIG. 8 alignement C).

L'alignement 1 (FIG. 2) peut être retrouvés en considérant tous les alignements contenus dans chaque noeud et chaque feuille de la structure alignée. L'alignement 2 (FIG. 3) est récupérable via l'alignement C (FIG. 8) et les alignements issus des autres noeuds.

5 Perspectives

En traitant ainsi un multidocument, nous souhaitons obtenir qu'une seule classe d'équivalence afin d'avoir un alignement sur toutes les langues du multidocument. Sur le corpus de 63 multidocuments traités, 27 multidocuments possèdent une seule classe d'équivalence, 16 sont divisés en deux classes (dont une représente 1 ou 2 documents). Les autres multidocuments sont divisés en plusieurs classes (< 6) ; une classe d'équivalence regroupe une majorité absolue des documents et les autres contiennent 1 ou 2 documents. Ces multidocuments sont composés de documents ayant des structures différentes : certains possèdent des annexes ou des titres là où les documents de la classe majoritaire n'en ont pas.



alignement A	et	it	nl
alignement B	1 1 2	a b	$\alpha\ \alpha\ \beta$
alignement C	3 1 1 1	c a a a a	$\gamma\ \alpha\ \alpha\ \alpha\ \alpha$

FIG. 8 – Exemple d'alignements pouvant être détectés automatiquement

Dans cette continuité, nous nous attacherons à établir des politiques de comparaisons de classes d'équivalence pour aligner les documents ayant des structures différentes. Une approche consiste à utiliser les cognates pour aligner les suites d'alinéas les contenant. Il est possible d'utiliser les langues proches (où la présence de cognates est forte) se situant dans des classes différentes pour les comparer. Une autre voie consiste à utiliser les longueurs des suites d'alinéas (de façon analogue à (Gale & Church, 1993)) pour envisager des alignements $m : n$ (où on pourra se demander si $m, n \geq 2$) entre ces suites (par exemple si des annexes ou des titres sont ajoutés). Enfin, nous pouvons calculer les distances d'éditions des arbres de structure.

Cette approche apporte des indices supplémentaires pour affiner les méthodes déjà existantes et avoir une nouvelle vue sur l'alignement.

Références

DEBILI F. & SAMMOUDA E. (1992). Appariements de phrases de textes bilingues français-anglais et français-arabes. In *Actes de COLING-92*, p. 528–524, Nantes : COLING-92.

GALE W. & CHURCH K. (1993). Indentifying word correspondences in parallel text. In *Fourth DARPA Speech and Natural Language Workshop*, p. 152–157.

GIGUET E. (2005). Multi-grained alignment of parallel texts with endogenous ressources. Borovets, Bulgaria : Modern Approaches in Translation Technologies Workshop.

KAY M. (1988). *Text-translation alignment*. Rapport interne, Xerox Palo Alto Research Center.

MELAMED D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, p. 107–130.

SIMARD M. (2000). Three languages are better than two. In J. VERONIS, Ed., *Parrallel Text Processing*.

SIMARD M., FOSTER G. & ISABELLE P. (1992). Using cognates to align sentences in bilingual corpora. In *proceedings of TMI-92*.

VÉRONIS J. (2000). Alignement de corpus parallèle. In J.-M. PIERREL, Ed., *Ingienerie des langues*, p. 151–171.

ZIMINA M. (2004). Topographie bitextuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles. Université de la Sorbonne Nouvelle.

Évaluation transparente de systèmes de questions-réponses : application au focus

Sarra EL AYARI
LIMSI-CNRS, BP 133, F-91403 ORSAY
Sarra.ElAyari@limsi.fr

Résumé. Les campagnes d'évaluation ne tiennent compte que des résultats finaux obtenus par les systèmes de recherche d'informations (RI). Nous nous situons dans une perspective d'évaluation transparente d'un système de questions-réponses, où le traitement d'une question se fait grâce à plusieurs composants séquentiels. Dans cet article, nous nous intéressons à l'étude de l'élément de la question qui porte l'information qui se trouvera dans la phrase réponse à proximité de la réponse elle-même : le focus. Nous définissons ce concept, l'appliquons au système de questions-réponses QALC, et démontrons l'utilité d'évaluations des composants afin d'augmenter la performance globale du système.

Abstract. Evaluation campaigns take into account only the final results obtained by information retrieval systems. Our perspective is that of the glass box evaluation of a question-answering system. The processing of a question is accomplished by a series of components. The purpose of this article is to study the element in the sentence which holds the key information. This element is to be found again in the sentence containing the answer next to the answer itself, and is called the focus. We will begin by defining this concept. We will then applied it to the QALC question answering system. Finally we will demonstrate the pertinence of using glass box evaluations to enhance the global performance such systems.

Mots-clés : système de questions-réponses, recherche d'information, évaluation, focus.

Keywords: question answering system, information retrieval, evaluation, focus.

1 Introduction

Les systèmes de recherche d'information (RI) ont vu apparaître de nouveaux types d'outils appelés systèmes de questions-réponses (SQR). C'est la prise en compte du besoin d'information précise de l'utilisateur qui a motivé l'émergence de tels systèmes.

Un SQR peut être opposé à un moteur de recherche sur Internet comme *Google* ou *Yahoo !* sur certains points bien précis. L'utilisateur saisit sa requête en langue naturelle (LN) et non sous la forme de mots clés. En aval, le système propose la ou les réponse(s) attendue(s) par l'utilisateur, et ne lui renvoie pas quelques milliers de documents qu'il doit parcourir manuellement.

Deux utilisations différentes s'esquissent clairement entre les deux types d'outils. Tandis que les moteurs de recherche permettent de récupérer des documents sur un thème général, les systèmes de questions-réponses sont utilisés pour trouver une information précise, qui tient en quelques

mots. Par exemple, la requête, qui est donc une question, « What is the FARC ? »[1] attend la réponse suivante : *the Revolutionary Armed Forces of Colombia*.

Les conférences organisées pour évaluer les systèmes de questions-réponses prennent uniquement en compte le résultat final obtenu. Or des traitements, des hypothèses sont instaurés à différentes étapes de la résolution des questions. Nous nous intéressons à une évaluation de type boîte transparente, qui veut évaluer les composants du système isolément. Cette évaluation est appliquée à l'étude d'un élément central pour l'extraction de la réponse : le focus. Sa reconnaissance intervient au moment de l'analyse de la question ; il constitue l'élément informationnel présent dans la question situé à proximité de la réponse dans la phrase réponse.

Nous présentons ce qu'est un système de questions-réponses (2), avant de nous focaliser sur le système du LIMSI : QALC dans une perspective d'évaluation de type « glass box » (boîte transparente). Après avoir présenté la dichotomie « black box » (boîte noire) / « glass box » (3), nous appliquerons ces principes à l'étude du focus (4).

2 Systèmes de questions-réponses

2.1 Architecture d'un système de questions-réponses

Un système de questions-réponses prend une question en entrée et doit fournir une réponse courte et précise à cette question. Nous travaillons sur le système *QALC* développé au LIMSI. QALC est conçu pour « traiter des questions factuelles ou encyclopédiques portant sur n'importe quel domaine » (Ferret *et al.*, 2001). En effet, ce système travaille en domaine ouvert, ce qui implique une certaine robustesse des processus employés. Pour ce faire, le système est composé de différents modules que nous allons expliciter. La chaîne de traitement est présentée sur le schéma 1[2].

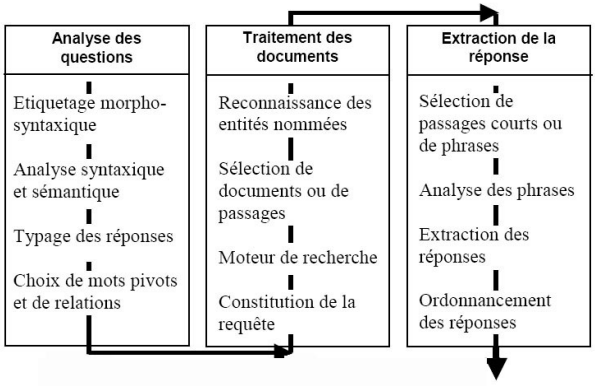


FIG. 1 – Chaîne de traitement de QALC

[1] Cette question est extraite du corpus CLEF 2005.
[2] Ce schéma est extrait de (Grau, 2004a).

Le traitement effectué par un SQR se fait en une séquence de trois étapes : l'analyse de la question, le traitement des documents sélectionnés par le moteur de recherche et l'extraction de la réponse dans les documents récupérés. Le traitement est ici plus fin que celui des moteurs de recherche : des traitements supplémentaires sont effectués en amont et en aval du moteur de recherche.

Il devient nécessaire de réaliser un travail plus fin sur la langue, qui fait appel au traitement automatique de la langue. Les informations principales dont on dispose se trouvent dans la question. Il s'agit d'en tirer le plus d'éléments pertinents possibles.

2.2 Le module d'analyse des questions

Le module d'analyse des questions permet de récupérer des informations essentielles pour identifier la réponse correspondant à la question. Nous présentons différents éléments qui importent pour le repérage de la réponse pour lesquels des traitements linguistiques peuvent être réalisés.

– Le type de réponse attendu, dans le cas où la réponse est une entité nommée (c'est-à-dire une unité d'un élément discursif qui fait référence à une personne, un lieu, une organisation, etc.). Ces entités sont repérées comme telles dans les documents extraits par le moteur de recherche. Elles sont déterminées par le pronom interrogatif utilisé ainsi que par des critères syntaxiques et sémantiques : *qui* indique que la réponse devra être une personne, *où* indique un lieu, *quand* indique une date, etc. S'il s'agit d'une question de type *Quel président français a été élu deux fois ?* le système repère une entité nommée de type personne.
– Le type sémantique de la réponse, lorsque celui-ci est explicite, est déterminé par un terme générique dans lequel est inclus le terme sur lequel la question porte. Pour *Quelle est la capitale du Togo ?* le type sémantique de la réponse sera lieu, par extension sémantique du terme *capitale*.
– L'objet de la question, à savoir le focus, qui correspond au mot de la question le plus important, défini comme étant celui que l'on doit retrouver dans la phrase réponse. Il s'agit du sujet de la question en quelque sorte.
– La catégorie de la question, déduite de son analyse syntaxique.
– Une extension sémantique de la question, qui consiste en la recherche de synonymes ou/et d'hyperonymes.

Par exemple, le traitement de la question *Qui a tué Henri IV ?* devra obtenir les informations suivantes :

- type attendu de la question : PERSONNE (entité nommée)
- objet de la question : Henri IV
- catégorie de la question : qui
- forme syntaxique de la question : PERSONNE VP[3] Focus
- forme syntaxique de la réponse :
  – REPONSE VP Focus pour ***<Réponse attendue>*** *a tué Henri IV*.
  – FOCUS VP REPONSE pour *Henri IV a été tué par* ***<Réponse attendue>***.
  – REPONSE nominalisation du verbe FOCUS pour ***<Réponse attendue>****, le tueur d'Henri IV*.
- extension sémantique : poignarder, assassiner, abattre

[3] VP est l'abréviation de verbe principal.

3 Évaluation fine des processus mis en place

3.1 Évaluation de type « black box » (boîte noire)

Les enjeux d'évaluation des systèmes de recherche d'informations ont pris une importance plus forte avec l'avènement du Web et le développement industriel des traitements textuels avec des infrastructures comme celle de la DARPA (Defense Advances Research Projects Agency), le NIST (National Institute of Standards and Technology) ou encore LDC (Linguistic Data Consortium) principalement aux Etats-Unis, et une prédominance de la langue anglaise.

En ce qui concerne les systèmes de questions-réponses, on distingue essentiellement deux campagnes internationales que sont TREC[4] (Text REtrieval Conference) et CLEF[5] (Cross Language Evaluation Forum), ainsi qu'une campagne française : EQUER[6] (Évaluation en Question Réponse). Nous pouvons également citer NTCIR, où l'évaluation porte uniquement sur des systèmes qui traitent les langues japonaise et chinoise.

La première campagne d'évaluation en questions-réponses a eu lieu en 1999 : il s'agit de la huitième édition de TREC, qui ne portait jusqu'alors que sur les systèmes de RI. Le déroulement de la tâche s'est complexifiée au fil des années. Si les premières campagnes proposaient 200 questions, pour lesquelles les participants pouvaient proposer cinq réponses par question sur des questions uniquement factuelles, il s'agit désormais de plus de 500 questions, qui peuvent être factuelles, mais qui peuvent également porter sur des listes, des questions définitionnelles et des scénarios (questions liées aux précédentes). Enfin, certaines questions n'ont pas de réponse dans les documents du corpus, et le système doit pouvoir l'indiquer. Une seule réponse courte est exigée, et celle-ci doit être accompagnée d'une justification (extrait de phrase permettant de valider la réponse proposée). De la même façon, la taille du corpus a augmenté de 528 000 documents en 1999 à 1 033 000 en 2005.

Ces campagnes permettent une évaluation globale des systèmes entre eux. Il s'agit de structures d'évaluation qui permettent de stimuler la recherche avec de nouveaux enjeux chaque année. Il devient alors possible de faire un constat mesuré de l'avancé technologique de ces systèmes chaque année. Un autre aspect important réside dans la mise au point d'étalons de référence pour évaluer ces systèmes avec des processus bien définis et des méthodes de comparaisons qui se veulent objectives.

Néanmoins, si la nécessité de ces campagnes d'évaluation n'est plus à démontrer, on peut tout de même avancer quelques réserves quant à leur évaluation car seul le résultat final obtenu par chacun des systèmes est pris en compte (classement selon les résultats obtenus). Elles ne permettent pas d'évaluer ce qui se passe à l'intérieur des systèmes de façon précise.

3.2 Évaluation de type « glass box » (boîte transparente)

Contrairement aux campagnes présentées, une évaluation de type boîte transparente donne un accès plus fin aux résultats produits par le système. Nous travaillons sur des systèmes composés de plusieurs modules qui s'enchaînent les uns après les autres (traitement de la question, des

[4]http ://trec.nist.gov/
[5]http ://clef.isti.cnr.it/
[6]http ://www.technolangue.net/article195.html/

documents puis extraction de la réponse). On voit alors la pertinence de telles évaluations : le résultat d'un module est pris en entrée par celui qui suit, et le résultat obtenu en aval dépend alors de la qualité des traitements effectués. Ces nouvelles formes d'évaluation deviennent nécessaires (Sparck Jones, 2001).

Il s'agit donc d'observer les résultats produits par les différents modules isolément, sans avoir à lancer le processus dans sa globalité, lequel ne sera pas forcément révélateur des problèmes existants. Une évaluation sélective et partitionnée permet de mesurer l'efficacité réelle de tel ou tel processus, afin de maximiser l'équilibre entre temps de traitement et efficacité. En effet, certains traitements peuvent être longs à s'exécuter sans pour autant être significatifs au niveau des résultats. C'est sur ce point qu'une évaluation interne permet d'améliorer un système. Évaluer de façon pointue permet de visualiser de façon précise ce qui n'est pas correct, mais aussi d'évaluer la « rentabilité » de certains traitements.

Plus précisément, dans le cadre qui nous intéresse, analyser les résultats obtenus par le système va permettre la redéfinition de certaines notions comme le focus, définitions qui seraient trop larges ou bien trop strictes par apport à l'utilisation qui en est fait. De plus, une mise en place d'un procédé d'évaluation transparente permet de modifier certains traitements, d'en ajouter ou bien d'en enlever et de tester la pertinence de ces modifications.

Nous allons appliquer cette méthode évaluative à l'étude de la notion de focus dans le module d'analyse de la phrase dans le système de question-réponse QALC.

4 Application à l'étude de la notion de focus

Le focus est un élément informationnel fort dans un énoncé (qu'il soit écrit ou oral), qui est utilisé pour aider à l'extraction de réponses dans QALC.

4.1 La notion de focus dans la littérature

La notion de focus prend ses origines dans la linguistique, et notamment en syntaxe et en phonologie. On parle de focus et d'opération de focalisation avec des niveaux d'analyse aussi variés que la syntaxe, la sémantique, la phonologie ou la phonétique. Ces disciplines considèrent que le focus est un élément informationnel important de la phrase. Il peut se manifester par une mise en valeur intonative à l'oral, et est l'objet mis en exergue dans les phrases clivées (*C'est cette poupée que je veux*).

Partant de ce constat, Wendy Lehnert a été la première à appliquer ce concept de focus à l'étude des questions pour les systèmes de questions-réponses (Lehnert, 1978). Elle définit alors le focus comme le concept de la question qui représente le besoin d'information exprimé par la question.

Plusieurs systèmes de questions-réponses ont intégré la reconnaissance du focus à leur traitement de la question, et en ont donné une définition.

– Pour (Ferret *et al.*, 2002) :
  « l'élément important de la question, celui qui devra se trouver à proximité de la réponse ».
– Pour (Plamondon *et al.*, 2002) :
  « une portion de la question qui doit obligatoirement figurer près du candidat-réponse [...].

Par exemple, le focus de la question *What was the monetary value of the Nobel Peace Prize in 1989 ?* serait Nobel Peace Prize car l'hypothèse est faite que la réponse correcte devrait se trouver la proximité de l'expression Nobel Peace Prize ou d'une expression sémantiquement apparentée ». Leur système est XR3 [7], premier système de question-réponse développé à l'Université de Montréal.

– Pour (Mendes & Moriceau, 2004) :
« l'élément le plus important de la question i.e. le focus ».

Ces trois définitions, qui viennent de différentes équipes de recherche, mettent en relief l'intérêt de la reconnaissance d'un terme qui doit se trouver dans la réponse : le focus. Elles montrent le lien syntaxique qui peut exister au sein de la phrase réponse entre le focus et la réponse à la question.

4.2 Une (re)définition du focus

4.2.1 Intérêt du focus

Le système QALC se situe dans une perspective d'approche robuste, où peu de connaissances sémantiques sont utilisées. De ce fait, nous essayons de déduire de façon automatique le plus de caractéristiques de la question. Le focus apparaît alors comme un élément important à repérer, nécessaire à la sélection des documents ainsi que pour l'extraction de la réponse, dont il se situe à proximité.

4.2.2 Définition du terme

Le focus est un terme pivot pour extraire la réponse attendue, qui doit apparaître à proximité de la réponse (Ferret *et al.*, 2002). Le système recherche le focus dans la phrase réponse, puis applique des patrons d'extraction par rapport à sa position dans la phrase réponse. Ce terme focus est un élément important pour l'extraction de la réponse courte attendue.

Dans notre approche, il constitue le plus souvent le sujet de la question. En effet, pour *When was the treaty on Conventional Forces in Europe signed ?* le focus est le sujet de la question : *treaty*. Pour des questions de type *Which EU conference adopted Agenda 2000 in Berlin ?* il s'agit alors du complément d'objet direct *Agenda 2000*. Considérer le focus comme le sujet ou l'objet d'un verbe sont des choix effectués afin d'implémenter cette notion de focus.

Dans notre expérience, le focus correspond à un groupe nominal présent dans la question, qui doit apparaître à proximité de la réponse. Nous différençons le focus du type général, qui renseigne sur le type de la réponse attendue. Cette différenciation est très claire si nous l'illustrons d'exemples :

– Which genes cause cancer ?
L'information recherchée est un type de gène. *Genes* constitue le type général de la question. Par contre, le focus, c'est-à-dire le terme autour duquel la réponse à la question s'articule, est *cancer*.

[7] Il s'agit de l'acronyme de eXtraction de Réponses Rapide et Robuste.

– Which US Army Division provided the paratroopers who took part in the invasion of Haiti ?
  Le focus repéré est *paratroopers* et *US Army Division* constitue ici le type général.
  Ce premier est un indice pour trouver la réponse, l'autre permet de renseigner sur la nature de l'information recherchée.

Mais cette définition du focus est-elle suffisante ? C'est ce que nous allons tenter de mesurer en analysant les résultats produits par notre système.

4.3 Observation des données

Nous effectuons dans cette étude une évaluation transparente du module d'analyse des questions en évaluant la reconnaissance ou non du focus et en observant en aval le nombre de réponses correctes obtenues pour les 200 questions dont nous disposons. Nous travaillons sur le corpus de CLEF 2005, campagne à laquelle le LIMSI a participé avec le système QALC. Nous disposons de 200 questions, réparties en 16 catégories.

Après avoir défini ce que nous considérons comme le focus, nous avons constitué une base de données afin d'observer en détail le traitement du focus effectué par le système de questions - réponses QALC.

Pour l'instant, le système reconnaît comme focus le premier groupe nominal rencontré qui est différent du type général. Il s'agit de faire une approximation de la notion de sujet.

Voici un exemple de question qui contient un type général et un focus : *In which year did the Islamic Revolution take place in Iran ?* où *year* constitue le type général et *Islamic Revolution* le focus. L'analyse de la question aboutit bien à la reconnaissance de ces deux éléments. La réponse extraite est correcte : *since the 1979 Islamic Revolution*. Cet exemple nous permet de légitimer la prise en compte de ces deux informations dans l'analyse des questions. Elles sont distinctes et nécessaires à la résolution des questions.

Nous avons effectué une étude de corpus manuelle afin d'observer les cas où la reconnaissance du focus pose problème au système.

4.3.1 Focus non repéré

Dans certains cas, le système ne repère pas le focus. Par exemple, *Which institution initiated the European youth campaign against racism ?* est une question pour laquelle aucun focus n'est spécifié. En effet, comme le premier groupe nominal repéré est le type général, le système ne recherche pas de focus. Il n'y a pas de réponse correcte trouvée pour cet exemple, alors que l'identification de *campaign* (pour se limiter à la tête du focus) aurait pu permettre au système une analyse des réponses plus complète.

4.3.2 Focus erroné

Certains focus identifiés ne correspondent pas à notre définition et génèrent des réponses erronées. C'est le cas pour *Which EU conference adopted Agenda 2000 in Berlin ?* où le système identifie *EU conference* comme focus, alors qu'il s'agit du type général. Or, le type général ne peut pas toujours être utilisé pour extraire des réponses. Il s'agit d'un terme générique - le plus

souvent un hyperonyme - sur lequel porte la question et qui n'est pas forcément présent tel quel dans la réponse.

Si le système attribue comme focus le type général d'une question, il ne pourra logiquement pas trouver de réponse correcte. Dans *Which Russian city is twinned with Glasgow ?* nous voyons bien que *Russian city* ne peut être un terme pivot autour duquel rechercher la réponse. La réponse attendue est une ville russe, mais la phrase réponse ne comportera pas forcément cette information.

Quand le système se trompe de focus, en le confondant avec le type général, il ne trouve pas de réponse.

4.3.3 Focus difficile à déterminer

D'autres questions posent problème quant à la définition même d'un focus. En effet, si l'on prend l'exemple *What newspaper was found in Kiev in 1994 ?* le type général est *newspaper* mais nous n'avons pas de terme focus. Certaines questions sont plus difficiles à traiter comme *According to which government did radioactivity from Chernobyl stop at the Franco-German border ?* Doit-on se focaliser sur *government*, *radioactivity* ou encore *Franco-German border* ? En fonction de la définition donnée du focus, *government* désigne le type attendu de la réponse : nous recherchons l'instance d'un gouvernement. Par contre, en ce qui concerne le focus, le terme doit être lié syntaxiquement à la formulation de la réponse. *Radioactivity* apparaît alors comme un bon candidat. Il s'agit là encore du sujet de la question, qui devra se trouver à proximité de la réponse attendue.

5 Conclusion

La perspective de cette première analyse est de regarder plus finement la définition du focus, de façon à maximiser les résultats obtenus. Il sera intéressant de voir comment formaliser le focus pour répondre aux difficultés soulevées dans cet article, afin que le système le retrouve automatiquement. Cela suppose d'affiner la définition du focus, et de modifier les patrons d'extraction de la réponse qui lui sont liés. A plus long terme, nous nous intéressons au développement d'une méthodologie d'évaluation transparente des systèmes de questions-réponses, de façon à affiner les traitements, et essentiellement à proposer des traitements différents selon les questions. Il s'agira de revoir la typologie effective des questions en fonction d'éléments comme la structure syntaxique de la réponse ou encore le type de focus. Surtout, nous pensons également qu'une étude fine des réponses obtenues ainsi que des réponses qu'il aurait fallu obtenir pourra nous permettre de valider ou encore une fois d'affiner notre définition du focus, ainsi que de mesurer la pertinence de cette information dans notre système.

Nous nous proposons de plus d'affiner notre définition du focus en fonction des réponses obtenues. Il serait intéressant de pouvoir par la suite réaliser une étude comparative avec un autre système de questions-réponses : RITEL[8] (Rosset *et al.*, 2006) afin de tester la validité de nos hypothèses.

En nous inscrivant dans une démarche d'évaluation transparente de systèmes de questions-

[8]Plus d'informations sur http ://ritel.limsi.fr/.

réponses, nous nous sommes intéressée à l'étude du focus, élément essentiel pour l'extraction de la réponse attendue de la question. Cette première observation nous a permis de voir que la redéfinition du concept utilisé permet d'affiner les résultats et de les améliorer. Cette démarche n'est pas possible lors d'évaluations de type boîte noire, or elle est plus que nécessaire pour l'amélioration des systèmes, en particulier des questions-réponses qui sont formés de plusieurs composants qui interagissent entre eux de manière séquentielle. Ce type d'évaluation apparaît bel et bien comme complémentaire aux évaluations de type boîte noire, afin d'améliorer finement les traitements effectués dans l'optique, par la suite, de l'obtention d'une meilleure performance globale.

Remerciements

Un énorme merci à Anne-Laure Ligozat pour son aide, ses encouragements et ses bons conseils. Grand merci aussi à Brigitte Grau et Benoît Habert pour leur aide et leurs relectures.

Références

BERTHELIN J.-B., GRAU B. & HURAULT-PLANTET M. (2001). Two levels of evaluation in a complex NL system. *Workshop on Evaluation for Language and Dialogue Systems*.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G. & JACQUEMIN C. (2001). Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. *Actes de TALN*.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., MONCEAUX L., ROBBA I. & VILNAT A. (2002). Recherche de la réponse fondée sur la reconnaissance du focus de la question. *Actes de TALN*.

GRAU B. (2004a). Evaluation des systèmes de question-réponse. In *Évaluation des systèmes de traitement de l'information*, chapter 3, p. 77–98. Hermès.

GRAU B. (2004b). Les systèmes de question-réponse. In *Méthodes avancées pour les systèmes de recherche d'informations*, chapter 10, p. 189–218. Hermès.

HARABIGIU S. & MOLDOVAN D. (2003). Question Answering. *Revue Computational Linguistics*.

HARABIGIU S., MOLDOVAN D., PASCA M. & SURDEANU M. (2003). Performance Issues and Error Analysis in an Open-Domain Question Answering System. *ACM Transactions on Informations Systems*.

LEHNERT W. (1978). *The Process of Question Answering : A Computer Simulation of Cognition*. John Wiley & Sons Inc.

MENDES S. & MORICEAU V. (2004). L'analyse des questions : intérêt pour la génération des réponses. *Workshop Question-Réponse*.

PLAMONDON L., KOSSEIM L. & LAPALME G. (2002). The quantum question answering system at trec-11. In E. M. VORHEES & D. K. HARMAN, Eds., *Proceedings of the Eleventh Text Retrieval Conference (TREC-2002)*, p. 750–757, Gaithersburg, Maryland : NIST.

ROSSET S., GALIBERT O., GABRIEL I. & MAX A. (2006). Interaction et recherche d'information : le projet RITEL. *Revue TAL*.

SPARCK JONES K. (2001). Automatic language and information processing : rethinking evaluation. In *Natural Language Engineering*, chapter 7, p. 1–18.

VOORHEES E. M. & HARMAN D. K. (2005). *TREC : Experiment and Evaluation in Information Retrieval*. MIT Press.

La segmentation thématique TextTiling comme indice pour le repérage de segments d'information évolutive dans un corpus de textes encyclopédiques

Marion LAIGNELET[1,2], Christophe PIMM[3]
[1] CLLE-ERSS – Université Toulouse 2 – Le Mirail, Toulouse
[2] Société INITIALES – Montpellier
[3] CLLE-ERSS – Université Toulouse 2 – Le Mirail, Toulouse
{marion.laignelet,christophe.pimm}@univ-tlse2.fr

Résumé. Nous faisons l'hypothèse que les bornes délimitées par la méthode statistique TextTiling peuvent servir d'indices qui, cumulées à des indices de nature linguistique, permettront de repérer automatiquement des segments d'informations évolutives. Ce travail est développé dans le cadre d'un projet industriel plus général dont le but est le repérage automatique de zones textuelles contenant de l'information potentiellement évolutive.

Abstract. Our hypothesis is that the TextTiling's boundaries can be considered as clues we can use with other linguistic features to automatically detect evolving information segments. This work is developed as part of an industrial project aiming to automatically detect textual zones containing potentially evolving information.

Mots-clés : segments d'information évolutive, segmentation, algorithme TextTiling.

Keywords: evolving information, segmentation, TextTiling algorithm.

1 Introduction

Un segment d'information évolutive (SEDIS-ε) est une portion de texte de longueur variable contenant de l'information dont la particularité est d'être susceptible d'évoluer dans le temps. Cette notion d'évolutivité de l'information a pris naissance dans un contexte industriel particulier, l'édition, et plus précisément dans le cadre d'une problématique de mise à jour éditoriale de documents encyclopédiques. L'étude présentée dans cet article est partie prenante d'un projet plus global[1] dont le but applicatif final est de proposer à des rédacteurs chargés de mettre à jour les articles encyclopédiques un outil facilitant cette tâche de mise à jour de l'information. Cet article vise à montrer que l'association de méthodes linguistiques avec des méthodes statistiques peut représenter un apport non négligeable lorsqu'il s'agit de rendre compte de phénomènes complexes. Plus précisément, nous présentons une

[1] Projet mené dans le cadre d'une thèse CIFRE, partenariat entre l'ERSS, Toulouse, et la société Initiales, Montpellier.

expérimentation dans laquelle nous exploitons la notion de rupture thématique issue du TextTiling (Hearst, 1994) avec l'idée que cela peut contribuer à l'identification des segments recherchés (les SEDIS-ε). Nous proposons ainsi un point de vue particulier sur la notion de segmentation des textes en optant pour une vision à gros grain de la segmentation, puisque d'un côté nous cherchons à décrire les SEDIS-ε, et puisque de l'autre, le TextTiling a pour but de segmenter les textes en fonction des thèmes qui sont abordés dans leurs parties. Notre hypothèse est que les ruptures thématiques fournies par cet algorithme peuvent devenir des indices pour aider au repérage des frontières initiales et/ou finales de certains types de SEDIS-ε. La première partie de cet article présente la notion de segmentation et son importance lorsqu'on travaille en linguistique du discours. Dans la seconde partie, nous présentons la notion de SEDIS-ε ainsi que l'apport d'une méthode telle que le TextTiling pour la description et le repérage de tels segments de discours. La troisième partie fait état du protocole expérimental que nous avons suivis. Enfin, nous présentons les résultats et proposons une discussion dans la dernière partie.

2 Segmentation

Ce projet peut être (entre autres) situé à la fois dans le sillon de la Recherche d'Information (et de la recherche intradocumentaire, par extension) parce que nous avons comme objectif de rechercher parmi la masse de textes encyclopédiques ceux pour lesquels une mise à jour est nécessaire, et dans celui de l'Extraction d'Information puisque nous souhaitons rechercher et extraire une information qui précisément présente la caractéristique d'être évolutive. Une des techniques commune à ces deux domaines est la segmentation[2].

2.1 Point de vue linguistique et cognitif

Nous suivons Péry-Woodley (2005) lorsqu'elle écrit : *« Toute structuration passe en effet par une segmentation, segmenter impliquant à la fois diviser et regrouper en fonction d'un critère organisationnel. Que l'on envisage l'organisation discursive en termes de structure d'information, de structuration thématique ou de relations de cohérence, la notion de segmentation est présente : recherches de critères de regroupement d'unités (en segments), identification de marques de rupture ou de discontinuité (entre segments), étude des relations (entre segments) qui les hiérarchisent et forment des segments de niveau supérieur. L'identification de segments à même de présenter une homogénéité sémantique et/ou de constituer des unités fonctionnelles est également au coeur des recherches en T.A.L. touchant au discours. »* (Péry-Woodley, 2005)

Il est important de souligner que, sous le terme de « segmentation », nous ne nous limitons pas à la notion de « segmentation thématique », laquelle s'appuie sur la notion de cohésion lexicale (*cf.* Halliday et Hasan, 1976), c'est-à-dire la répétition des mots comme indicateur d'homogénéité thématique. De nombreux travaux, qu'ils soient menés en linguistique ou en psycholinguistique, cherchent à rendre compte de ces processus de construction de la cohérence que ce soit à travers la délimitation de segments de discours ou de la description des relations entre ces segments (*cf.* RST, SDRT, segmentation thématique, etc.). La segmentation automatique apparaît alors comme une technique incontournable développée relativement à divers objectifs applicatifs : pour certains l'objectif visé est celui du résumé automatique (*cf.* Minel (2002), Marcu (2000), etc.) ; pour d'autres il s'agit de développer des

[2] La segmentation peut également être vue comme une tâche à part entière, parti que nous n'adoptons pas dans ce cadre. Nous l'envisageons comme une étape nécessaire à un grand nombre de traitements sur les textes.

systèmes de recherche/sélection d'information importante (*cf.* Rossi et Bert-Erboul (1991), etc.) ou encore des outils d'aide à la navigation (*cf.* Jackiewick et Minel (2003), etc.). Dans tous ces cas, la tâche de segmentation est envisagée par rapport à une application précise, un point de vue sur les textes (*cf.* Hernandez, 2004). C'est également la position que nous adoptons : il nous paraît essentiel et incontournable de définir les segments que nous recherchons, les SEDIS-ε, ainsi que les méthodes et techniques mises en œuvre pour y aboutir relativement à l'objectif applicatif de départ, à savoir la mise en place d'un outil d'aide à la mise à jour de l'information dans des documents encyclopédiques.

2.2 Segmentation thématique et TextTiling

La segmentation thématique est un domaine riche pour lequel il y a de plus en plus de travaux. Généralement, sont développées des approches statistiques même si la tendance aujourd'hui consiste à les combiner à des analyses linguistiques souvent sommaires. Nous ne présenterons ici qu'une approche, le TextTiling de Hearst (1997), qui reste encore aujourd'hui la méthode la plus utilisée pour ce type de segmentation. Selon Hearst (1997), cette tâche de segmentation peut potentiellement s'intégrer dans des applications d'extraction d'information ou de résumé automatique. Elle décrit un algorithme à deux tâches principales : d'un côté, l'identification des *subtopic segments* et de l'autre le repérage des *subtopic shifts*. Elle travaille sur les paragraphes ou sur des ensembles comprenant plusieurs paragraphes. Au final, le *TextTiling Algorithm* a pour objectif de segmenter le texte en plusieurs blocs contigus, qui ne se chevauchent pas et qui sont cohérent thématiquement. Des scores sont ensuite attribués à ces blocs de texte, et c'est l'attribution de ces scores qui participe à la segmentation. Nous développerons dans la 3^{e} partie les diverses étapes de la segmentation telle qu'elle est envisagée dans le cadre du TextTiling mais également telle que nous l'utilisons.

3 Les SEDIS-ε dans le cadre d'un projet industriel

3.1 Présentation et Définitions

Nous définissons un SEDIS-ε comme un segment textuel susceptible de contenir une ou plusieurs informations présentant cette particularité de pouvoir évoluer dans le temps et/ou qui relativement à des besoins éditoriaux nécessiterai(en)t d'être réactualisée(s) (*cf.* Laignelet, 2006a, 2006b, 2006c). Une double distinction notionnelle nous permet de rendre compte partiellement de la complexité de la notion de « mise à jour » qui peut être envisagée tant du point de vue de la tâche réelle à laquelle elle fait référence que du point de vue du linguiste dont l'objectif est de décrire l'objet textuel auquel il réfère. Nous faisons donc une distinction sur deux plans. Le premier plan concerne la nature de l'information à mettre à jour, laquelle peut être à strictement parler une mise à jour ou bien une réactualisation. Dans le cas de la **mise à jour**, l'information n'est plus vraie ou ne s'est pas vérifiée (c'est souvent le cas lorsque l'auteur fait des prédictions sur un fait ou un événement). Dans l'exemple suivant, l'information *« Il n'existe pas à l'heure actuelle de vaccin contre le sida. »* ainsi que ce qui suit n'est potentiellement plus vrai au moment de lecture/réédition ou alors, étant donné un possible caractère prédictif, on est en droit de se demander si elle s'est ou non vérifiée.

La découverte du virus a permis la mise au point d'une méthode de dépistage [...]. On peut ainsi savoir qu'une personne est infectée longtemps avant que la maladie ne se déclare. Il n'existe pas à l'heure actuelle de vaccin contre le sida. Si les thérapies actuelles permettent d'améliorer sensiblement la durée et les conditions de vie du malade, aucune n'est capable d'éliminer le virus.

Figure 1 : Exemple d'une mise à jour

Dans le cas d'une **réactualisation,** les segments contiennent une information qui restera vraie dans l'absolu mais, en vue d'une ré-édition et d'une diffusion, les événements et dates associés doivent être modifiés pour faire référence à un moment plus proche du moment de lecture/réédition. Dans l'exemple qui suit, la valeur chiffrée *« 160 millions »* associée à *« en 2002 »* reste vraie, qu'on lise la fiche en 2003 ou en 2007. Cependant, il est fortement souhaitable de fournir de nouvelles informations et notamment de donner les chiffres pour l'année la plus proche de la date de réédition de la fiche.

L'organisation mondiale de la santé (OMS) estime, en effet, à 160 millions le nombre annuel de nouveaux cas dans le monde en 2002.

Figure 2 : Exemple d'une réactualisation

Le second plan fonde la distinction à un niveau plus textuel et oppose le SEDIS-ε minimal au segment d'interprétation. Les SEDIS-ε sont ainsi envisagés comme des segments textuels à granularité variable, ce qui nous permet de répondre au mieux aux exigences industrielles : en effet, il semble préférable pour le rédacteur chargé de la mise à jour d'avoir accès à la fois aux expressions locales à mettre à jour à proprement parler, et en même temps de bénéficier d'un contexte textuel suffisamment large et plus global pour être en mesure d'interpréter et de cibler rapidement ce qui nécessite une mise à jour. Dans la figure 3, nous pouvons voir un certain nombre de SEDIS-ε minimaux. Ils apparaissent dans cet exemple dans les petits cadres (ovales et rectangulaires). Leur taille varie du mot au syntagme et ils peuvent être de diverses natures. Dans certains cas, ils se confondent avec la notion d'indice : c'est le cas notamment des dates et des valeurs chiffrées. L'ensemble de l'extrait correspond à la notion de segment d'interprétation : c'est un segment discursif plus long que les précédents ; il s'agit ici de l'exemple en entier. La taille minimale de ce type de segment est la phrase mais ils peuvent aussi couvrir un ou plusieurs paragraphes, voire la partie entière.

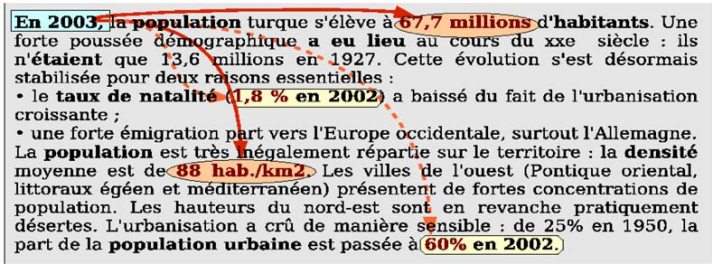


Figure 3 : Exemple présentant un cadre temporel ouvrant un segment d'interprétation

Dans cet exemple, le segment d'interprétation s'ouvre sur un introducteur de cadre temporel (Charolles, 1997). L'intérêt de considérer l'IC temporel *« En 2003 »* (dans le premier encadré) est que le critère sémantique (la référence temporelle *« 2003 »*) qu'il véhicule est valable pour

l'ensemble du segment donné. Ainsi, les deux valeurs chiffrées dans les ovales ont une relation (temporelle) à travers l'expression *« En 2003 »*. Les deux éléments dans les encadrés arrondis sont également des informations à mettre à jour du fait de leur proximité temporelle. Dans ce segment, il est important de noter que toutes les informations contenues ne sont pas à mettre à jour, par exemple *« Une forte poussée démographique a eu lieu au cours du XXe siècle [...] »,* pour lesquelles une référence temporelle différente est explicitement signalée. Nous définissons un segment d'interprétation comme un segment textuel de longueur indéterminée, présentant une certaine homogénéité sémantique (temporelle, spatiale, etc.), pouvant contenir des segments ne nécessitant pas de mise à jour et qui contient des SEDIS-ε minimaux et/ou des indices.

3.2 Repérage des SEDIS-ε : l'apport d'une méthode telle que le TextTiling pour le repérage des segments d'interprétation

Des travaux antérieurs menés dans une optique de segmentation ont cherché à combiner les marques de natures différentes, soit statistique et linguistique. Dans le cadre du projet REGAL, les auteurs cherchent à construire une structure thématique des textes afin de pouvoir soutenir une navigation intra-document dans le cadre de systèmes de résumé dynamique. Dans cette optique, Hernandez (2004) propose de combiner une analyse par segmentation lexicale avec un repérage de marques linguistiques. Cette étude nous semble très intéressante, à la fois de par la méthode utilisée mais également à travers les conclusions qu'elle apporte. Ainsi, selon Hernandez (2004 : 184), *« la cohésion lexicale apporte la robustesse au système en lui permettant de produire des résultats relevant de tout domaine [...]. Les marques linguistiques quant à elles apportent la finesse en permettant de repérer avec précision les bornes de segments, plus souvent la borne initiale d'ailleurs, la délimitation de la borne finale étant souvent très difficile voire impossible »*. Hernandez conclut en disant que *« la combinaison d'une analyse automatique par cohésion lexicale et d'un repérage des cadres apparaît comme triplement profitable : elle permet dans certains cas d'affiner l'ajustement des marques de segmentation automatique, dans d'autres de fournir un indice supplémentaire de fermeture de cadre ; enfin elle met en lumière un point important concernant la dimension lexicale des cadres de discours : il semble en effet que ceux-ci présentent une cohésion lexicale forte chaque fois qu'ils jouent un rôle procédural dans la classification des données transmises au lecteur. »* Dans le cadre de notre projet, nous cherchons à mettre en place une expérimentation telle que celle qui a été faite par Hernandez tout en étant conscient que notre objectif applicatif est bien différent. Ainsi, nous supposons que le repérage des SEDIS-ε peut être automatisé à travers la prise en compte d'indices linguistiques et discursifs, lesquels, bien qu'ils aient une fonction précise dans la langue, peuvent également permettre l'interprétation d'un segment comme étant de nature évolutive. S'ils sont considérés isolément, ces indices ne sont cependant pas suffisants pour répondre à cet objectif de repérage automatique des SEDIS-ε (Laignelet, 2006a) ; en revanche, envisagés en termes de configurations, *i.e.* en prenant en compte des combinaisons d'indices, il est tout à fait pertinent de penser que cette tâche peut être automatisée. En plus de la combinaison d'indices linguistiques et discursifs – et c'est l'objet de cet article – nous souhaitons analyser l'impact d'une analyse statistique de type TextTiling. Dans notre cas, et au stade de notre étude, nous souhaitons observer si les ruptures thématiques engendrées par une telle segmentation peuvent elles-mêmes devenir des indices pour le repérage des frontières initiales et/ou finales des SEDIS-ε de type « segments d'interprétation ».

4 Protocole expérimental

4.1 Annotation manuelle

Nous travaillons actuellement sur un corpus constitué de 92 textes de type encyclopédique. Il s'agit de fiches encyclopédiques éditées et accessibles sur le marché de l'édition (propriété des Editions Atlas). *A priori* du point de vue du type de texte (*cf.* terminologie de Biber), ce corpus est homogène. Le trait distinctif entre ces textes relève de la catégorisation en genre et plus précisément du domaine de connaissance auquel chacune des fiches appartient. Nous insistons sur ce point parce que nous supposons l'importance de cette distinction par domaine pour les résultats[3]. Ce corpus constitue une base de 80 000 mots, au format XML. Sur le total des 92 fiches, nous avons procédé à l'annotation manuelle de 38 d'entre elles. Cette tâche d'annotation manuelle a consisté à marquer à l'aide de balises XML les frontières des segments qui sont potentiellement des segments contenant de l'information à mettre à jour. Elle met en évidence la présence de 630 SEDIS-ε dans les 38 fiches parcourues. Par choix méthodologique[4], les SEDIS-ε sont de longueur égale à l'unité phrase (*i.e.* qui commence par une majuscule et se termine par une marque de ponctuation).

4.2 Outil : LinguaStream

Même si ce n'est pas l'objectif central visé par article, il est important de préciser que l'ensemble des marqueurs de surface[5] sur lesquels nous travaillons sont repérés de manière automatique à l'aide de la plateforme LinguaStream (*cf.* Widlöcher et Bilhaut, 2005). LinguaStream est une plate-forme générique pour le traitement automatique des langues qui permet d'effectuer des traitements et des analyses de types et de niveaux linguistiques variés (morphologique, syntaxique, sémantique, discursif ou encore statistique) sur des corpus en XML : il offre la possibilité d'utiliser différents langages en fonction de ce qu'on veut faire : des lexiques, des grammaires Prolog, des expressions régulières, des macro-expressions régulières, des programmes groovy, etc. . Travaillant sur des objets « mouvants » (qui peuvent aller de la taille d'une phrase à la taille d'une partie entière) cet outil est particulièrement pertinent de par les diverses possibilités de visualisation qu'il offre. Concernant cette étude, LinguaStream nous permet de travailler directement sur notre corpus annoté manuellement des SEDIS-ε, d'y adjoindre un découpage issu d'une segmentation TextTiling et de comparer, observer et analyser aisément les deux types de segmentation obtenus.

4.3 Détail des étapes de la segmentation par le TextTiling et adaptation à notre étude

L'algorithme du TextTiling permet une segmentation des textes fondée sur la notion de changement thématique. L'hypothèse de Hearst (in Hernandez, 2004 : 191, *note n°86*) est que *« un ensemble d'items lexicaux est utilisé pendant la discussion d'un sous-thème, et quand le*

---

[3] Huit domaines différents sont représentés : géographie (14), médecine & santé (13), sciences & techniques (10), société (8), sport (8), histoire (17), art & littérature (12) et faune & flore (7). Nous ne traitons pour cette étude que les quatre premiers domaines cités.

[4] Cela nous permet notamment de pouvoir nous baser sur des unités homogènes (la phrase) lors de l'évaluation de nos programmes.

[5] des adverbes de temps, des syntagmes nominaux de temps, des superlatifs, des noms propres, des sigles, des superlatifs (Laignelet, 2006a, 2006b, 2006c)

sous-thème change, une proportion significative du vocabulaire change aussi ». D'une manière générale, cet algorithme consiste à comparer des paires de passages successifs après pondération des mots de chacun des passages en fonction de critères de distribution et de co-occurrence lexicale. Après une première étape de tokenisation et d'étiquetage pour supprimer les mots vides susceptibles de parasiter les traitements ultérieurs, Hearst pose une taille fixe et arbitraire des passages de texte à comparer : elle définit les notions de pseudo-phrase (*token sequence*) et de pseudo-paragraphes (*block*). Ces deux éléments présentent la particularité d'être de taille homogène, les pseudo-paragraphes comptant un nombre donné de pseudo-phrases. L'auteur explique qu'elle a choisi de ne pas utiliser les phrases comme unités car la phrase est un segment dont la définition pose problème et de plus, le fait de comparer des segments de même taille rend cette comparaison plus aisée. Une fois les blocs définis, des scores de similarité vont être calculés entre chaque paire de blocs adjacents (cf. cosinus, coefficient de Dice, Jacquard, etc.) basés sur la fréquence des tokens de chaque bloc. Dans une dernière phase, les frontières des segments thématiques sont détectées par comparaison des différences scores qui ont été attribués à chacun des segments lors de l'étape précédente. Dans le cadre de notre expérimentation, nous avons fait le choix de travailler sur des segments dont la taille fait trois phrases « réelles »[6]. Nous verrons que ce choix n'est pas sans poser de problème dans des textes où la mise en forme matérielle est très riche. Nous avons exclu de considérer l'unité paragraphe du fait de la particularité de notre corpus : en effet, s'agissant de fiches encyclopédiques grand public et donc à fort impact visuel, les personnes chargées de la mise en page de ces fiches ne respectent pas ou peu la signification du saut de ligne. Nous souhaitions donc une méthode nous permettant à la fois de descendre en dessous du grain paragraphe, et de traiter des unités plus ou moins homogène en taille. Par ailleurs, nous ne prenons pas non plus en compte les titres dans les calculs de similarité. Nous avons utilisé le coefficient de Dice tout en faisant varier le seuil en fonction des thématiques des textes : pour les fiches *géographie*, le seuil est de 0,09 ; pour les fiches *médecine & santé*, il est également de 0,09 ; pour les fiches *sciences & techniques*, il est de 0,15 ; enfin pour les fiches *société*, il est de 0,17. Cette variation a été mise en évidence après divers tests effectués sur notre corpus ; ces différences reflètent ainsi les variations que l'on peut observer entre types et genre de textes.

5 Premiers résultats

Une fois que les différents traitements que nous venons de présenter ont été effectués sur notre corpus, nous avons observé et compté le nombre de fois où, d'un côté une balise ouvrante de SEDIS-ε correspond à une balise ouvrante de segment thématique TextTiling (« ouverture stricte »), et de l'autre une balise fermante de SEDIS-ε correspond à une balise fermante de segment thématique TextTiling (« fermeture stricte »). De plus, du fait que les blocs aléatoires utilisés pour faire les calculs TextTiling peuvent avoir des frontières un peu n'importe où, nous avons également observé et compté les cas où il y avait une phrase de décalage entre les deux types de borne. Le tableau suivant récapitule ces données :

[6] Par « réelles » nous entendons des phrases grammaticales qui commencent par une majuscule et se terminent par une ponctuation forte. Cela ne correspond donc pas à la notion de *pseudo-phrase* définie par Hearst.

	valeurs réelles		Pourcentage	
nombre de SEDIS-ε	630		100	
nombre de « fermetures strictes »	141	200	22,38 %	31,75 %
nombre de fermetures +/-1 phrase	59		9,36 %	
nombre de « ouvertures strictes »	114	172	18,09 %	27,30 %
nombre d'ouvertures +/-1 phrase	58		9,20 %	

Tableau 1 : Comparaison des frontières ouvrantes et fermantes des SEDIS-ε et des segments thématiques

Ces résultats nous encouragent fortement à considérer les ruptures thématiques issues d'une segmentation TextTiling comme des indices nous permettant d'automatiser le repérage de SEDIS-ε autant pour le repérage de leur borne initiale que pour celui de leur borne finale : en effet, 31,75 % des frontières finales de SEDIS-ε apparaissent simultanément avec une fin de segment thématique, et dans 27,30 % des cas, les frontières ouvrantes de SEDIS-ε et de segment thématique sont co-occurrents. La figure 4 montre un cas où l'on peut observer une co-occurrence à la fois des bornes initiales des deux types de segments et de leurs bornes finales.

<SEDIS-ε> <HEARST> En 2000, la Chine avait une production de pêche de capture estimée à 17 millions de tonnes. La France occupe le quatrième rang en Europe avec une production annuelle d'environ 600000 tonnes, poissons, crustacées et mollusques réunis. Les sources rapportées par la FAO (Food Agricultural Organization) font état d'une croissance du commerce halieutique international de 4 % par an, soit un montant évalué 55,2 milliards de dollar en l'an 2000. </HEARST> <SEDIS-ε>

Figure 4 : Exemple de co-occurrence entre SEDIS-ε et segment thématique

Ce qui est intéressant dans cet exemple 4, c'est qu'il montre également que d'autres indices peuvent être pris en compte comme les introducteurs de cadres temporels (ici *« En 2000 »*) (Charolles, 1997). Cela rejoint les conclusions de Hernandez (2004 :194) sur la relation entre cadre de discours et cohésion lexicale. Il apparaît donc indispensable de considérer la segmentation TextTiling non pas de manière isolée ou suffisante en elle-même, mais comme partie prenante de configurations d'indices de natures différentes (linguistiques et discursives). Mais les résultats que nous donnons dans le tableau 1 ne permettent pas de rendre compte d'un certain nombre de limites liées à la fois à notre méthode mais également liées à des difficultés inhérentes au type de corpus sur lequel nous travaillons. Nous venons de faire la remarque suivant laquelle il est nécessaire de considérer indices linguistiques et indice statistiques de manière complémentaire. En fait, la prise en considération de certains indices discursifs devrait être effective au moment de la délimitation des segments aléatoires : ceci permettrait par exemple dans l'extrait de la figure 5 de faire débuter le segment au moment où un introducteur de cadre est également présent. On peut alors supposer que la balise ouvrante <HEARST> se situerait simultanément avec l'ouverture d'un cadre temporel et dans ce cas précis avec celle d'un SEDIS-ε.

[...] Actuellement, la place très importante de la France au sein de l'ONU lui impose de répondre aux menaces régionales. <HEARST> Elle participe aux opérations de maintien de la paix sous l'égide de l'ONU. Des casques bleus français sont ou ont été présents en République Centrafricaine, en ex-Yougoslavie, à Jérusalem [...] </HEARST> [...]

Figure 5 : Limites de notre méthode : prendre en compte des indices discursifs comme les introducteurs de cadre de discours

Par ailleurs, nous avons déjà souligné le fait que les textes sur lesquels nous travaillons sont des textes dans lesquels la performance visuelle est prégnante. Ceci entraîne donc une mise en forme matérielle très riche que la segmentation aléatoire ne prend pas en considération. Concernant l'algorithme TextTiling, Hearst l'a évalué sur des textes descriptifs (*« expository texts »*) ce qui est très différent du type encyclopédique sur lequel nous avons travaillé.

 En revanche, certaines sectes [...], que les Renseignements généraux ont listées : - la déstabilisation mentale ; - le caractère exorbitant des exigences financières ; - la rupture induite avec l'environnement ;   - les atteintes à l'intégrité physique [...] ; - l'embrigadement des enfants ; - le discours plus ou moins anti-social ;   - les troubles à l'ordre public ; - l'importance des démêlés judiciaires[...]

Figure 6 : Limites de notre méthode : prendre en compte la mise en forme matérielle

Dans cet exemple, nous pouvons voir que la segmentation aléatoire se place n'importe comment au sein de l'énumération : nous proposons donc de modifier cette phase essentielle de segmentation du TextTiling en prenant en considération des indices relevant de la mise en forme matérielle. Il est ainsi pour les énumérations mais également pour les parties titrées courtes (de niveau 3).

6 Conclusion

L'hypothèse selon laquelle les frontières de segments thématiques peuvent servir d'indices contribuant au repérage des segments d'information évolutive semble se confirmer même s'il s'avère nécessaire d'approfondir les expérimentations présentées. L'objectif de ce travail n'est pas de montrer que les segments thématiques correspondent exactement aux SEDIS-ε mais il consiste à observer la possibilité d'exploiter les bornes initiales et/ou finales de tels segments conjointement à d'autres indices de natures diverses (adverbiaux de temps, superlatifs, temps verbaux, etc.) pour le repérage des SEDIS-ε. Ainsi, nous envisageons notamment d'adapter l'algorithme TextTiling et plus précisément la phase de découpage en pseudo-blocs avec des techniques de segmentation de nature linguistique telles que l'encadrement du discours ou encore les titres pour améliorer nos résultats et notre objectif général.

Références

CHAROLLES M. (1997). L'Encadrement du Discours, Univers, Champs, Domaine et Espaces. *Cahiers de Recherche linguistique.*

HALLIDAY M., HASAN R. (1976). *Cohesion in English.* Longman Group Limited, London.

HEARST M. (1994). Multi-paragraph segmentation of expository texts. Proceedings of *the 32th Annual Meeting of the Association for Computational Linguistics*.

HERNANDEZ N. (2004). *Description et Détection Automatique de Structures de Texte.* Thèse de doctorat, Université de Paris XI.

JACKIEWICZ A., MINEL J.-L. (2003). L'identification des structures discursives engendrées par les cadres organisationnels. *Actes de la 10ème conférence sur le traitement automatique des langues naturelles, TALN.* Batz-sur-mer, France.

LAIGNELET M. (2006a). Repérage de segments d'information évolutive dans des documents de type encyclopédique. *Actes de la 13ème conférence jeune chercheur sur le traitement automatique des langues naturelles, RECITAL.* Presses Universitaires de Louvain, Louvain, Belgique.

LAIGNELET M. (2006b). Analyse discursive pour le repérage de segments d'information évolutive. *74ème Congrès de l'ACFAS, Description Linguistique pour le Traitemant Automatique du Français (DLTAF-ACFAS).* 16-18 mai 2006, Montréal, Canada.

LAIGNELET M. (2006c). Les titres et les introducteurs de cadre come indices pour le repérage de segments d'information évolutive. *Actes du Colloque International Discours et Document (ISDD'06)*, Presses Universitaires de Caen, France.

MARCU D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*, The MIT Press.

MINEL J.-L. (2002). *Filtrage sémantique, du résumé automatique à la fouille de textes*, Hermès.

PERY-WOODLEY M.-P. (2005). *Discours, corpus, traitements automatiques.* Hermès.

ROSSI J., BERT-ERBOUL A. (1991). Sélection des informations importantes et compréhension de textes. *Psychologie Française*.

WIDLÖCHER A., BILHAUT F. (2005). La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus, Actes de la *12e Conférence Traitement Automatique du Langage Naturel (TALN).* Dourdan, France.

Annotation des disfluences dans les corpus oraux

Marie PIU, Rémi BOVE
Équipe DELIC – Université de Provence
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1
remi.bove@up.univ-mrs.fr, piumarie@yahoo.fr

Résumé. Les disfluences (répétitions, amorces, autocorrections, constructions inachevées, etc.) inhérentes à toute production orale spontanée constituent une réelle difficulté en termes d'annotation. En effet, l'annotation de ces phénomènes se révèle difficilement automatisable dans la mesure où leur étude réclame un jugement éminemment interprétatif. Dans cet article, nous présentons une méthodologie applicable à l'annotation des disfluences (ou « phénomènes de production ») que l'on rencontre fréquemment dans les corpus oraux. Le fait de constituer un tel corpus de données annotées, permet non seulement de représenter certains aspects pertinents de l'oral (de manière à servir de base aux observations et aux comparaisons avec d'autres données) mais aussi d'améliorer in fine le traitement automatique de l'oral (notamment l'analyse syntaxique automatique).

Abstract. Disfluencies (repeats, word-fragments, self-repairs, aborted constructs, etc.) inherent in any spontaneous speech production constitute a real difficulty in terms of annotation. Indeed, the annotation of these phenomena seems not easily automatizable, because their study needs an interpretative judgement. In this paper, we present a methodology for the annotation of disfluencies (also named "production phenomena") which frequently occur in speech corpora. Constituting such data allows not only to represent some relevant aspects of speech productions (so as to be a basis for observations and comparisons with other data), but also to improve automatic speech processing (particularly for parsing).

Mots-clés : corpus oraux, annotation, disfluences, prosodie, XML.

Keywords: speech corpora, annotation, disfluencies, prosody, XML.

1 Introduction

À l'heure actuelle, les études linguistiques qui basent leurs descriptions sur de vastes corpus électroniques gagnent sans cesse du terrain et les outils d'analyse automatique de plus en plus performants se multiplient. Malgré cela, l'analyse automatisée de l'oral reste marginale car on

ne dispose que de très peu de corpus oraux[1]. La constitution et l'annotation de corpus oraux représentent un enjeu de première importance en vue d'applications telles que reconnaissance vocale, apprentissage des langues, etc.

Par ce travail, nous proposons de constituer des données de référence annotées pour l'oral par le biais d'un schéma d'annotation et d'un formalisme adaptés. L'objectif de ce travail est double : il permet d'une part, d'obtenir des données de référence sur l'oral et d'autre part, de faciliter l'exploitation informatique de ces mêmes données. Pour mener à bien ce projet, nous avons jugé nécessaire de fonder notre analyse sur un cadre théorique existant qui traite des phénomènes de l'oral. En effet, pour garantir sa cohérence, il est indispensable que ce schéma d'annotation soit en adéquation avec un modèle d'analyse de l'oral préalablement défini. Nous nous sommes donc inspirés du modèle d'analyse de « la mise en grille » proposé par (Blanche-Benveniste, 1987), pour annoter les phénomènes de production dans notre corpus.

À notre sens, ce travail doit aussi lier analyse qualitative et quantitative pour rendre compte des multiples aspects du langage que l'on peut observer dans les situations d'oral. L'analyse qualitative se traduit par la mise en place d'un schéma d'annotation et par l'enrichissement des données. Ensuite, les informations quantitatives recueillies permettent de mieux connaître les mécanismes de la langue orale par les informations ponctuelles qui s'en dégagent (fréquence, répartition des phénomènes) mais aussi par les informations de structures (patrons récurrents, contexte étudié).

2 Corpus d'étude et phénomènes étudiés

Le corpus à partir duquel nous avons procédé à l'annotation des disfluences est une sous-partie du Corpus de Référence du Français Parlé (CRFP) constitué par l'équipe DELIC (Description Linguistique Informatisée sur Corpus) et spécialement choisie pour son caractère monologique et son hétérogénéité situationnelle.

2.1 Corpus de Référence du Français Parlé

Cette sous-partie du CRFP se compose de dix enregistrements (environ 53 minutes de parole, soit plus de 8000 mots) faisant intervenir cinq hommes et cinq femmes. La transcription orthographique a été effectuée entièrement à la main par des experts linguistes avec un cycle de réécoute/validation extrêmement strict. Les conventions de transcription adoptées (cf. Équipe DELIC, 2004) ne contiennent aucun trucage orthographique (du type p'tit, y'a, etc.) ni aucune ponctuation, suivant la tradition de l'équipe. Par ailleurs, un certain nombre de phénomènes de production à l'oral ont été transcrits (sans être annotés spécifiquement) : les répétitions, les amorces, les euh d'hésitation, les allongements, les pauses, les accents ainsi que les mouvements intonatifs majeurs.

[1] On entend généralement par « corpus oraux » les annotations (sous forme de transcriptions orthographiques) et les enregistrements (fichiers sons et/ou vidéo)

2.2 Phénomènes étudiés

Nous avons choisi de nous intéresser plus particulièrement à l'annotation des phénomènes de production à l'oral. Voici les phénomènes qui ont constitué notre objet d'étude et pour lesquels nous avons mis en place un schéma d'annotation :

- Les **répétitions** : répétition d'un ou plusieurs mots ou reprise à l'identique d'une syllabe, d'un mot ou d'une amorce de mot, de plusieurs syllabes ou de plusieurs mots, sans aucune valeur sémantique (Candéa, 2000).

  (1) *on entreposait **les: les: les** huiles ↗ (CRFP)*

- Les **autocorrections** : substitution d'un mot ou d'une série de mots à d'autres afin de modifier ou corriger une partie de l'énoncé (Kurdi, 2003).

  (2) *à cette époque **j'avais j'étais** en maîtrise il me restait le mémoire à faire ↗ (CRFP)*

- Les **amorces** : interruption de morphème en cours d'énonciation (Pallaud, 2002)

  (3) *donc je suis restée trois mois en **c-** en camping à peu près hein (CRFP)*

- Les **inachèvements** : énoncés auxquels il manque un ou plusieurs éléments pour qu'ils soient grammaticalement bien formé et interprétable sémantiquement (Kurdi, 2003).

  (4) *j'étais au bord de la mer c'était super ça été <u>un moment de:</u>*

- Les **disfluences combinées** : association simultanée d'au moins deux des phénomènes présentés ci-dessus.

  (5) *je voyais p- il y av- j'avais pas d'autre so- enfin j'avais pas d'autre solution ↘*

3 Cadre d'analyse des disfluences

Pour mener à bien notre projet et pour garantir la cohérence de notre méthode d'annotation, nous avons jugé nécessaire d'appuyer notre analyse sur un cadre théorique existant.

L'une des approches les plus répandues concernant la modélisation des disfluences est celle proposée par Shriberg (1994). L'auteur décrit l'organisation interne des disfluences en un ensemble d'espaces distincts délimitant les étapes de la production orale. Le *reparandum* (RM) correspond à la partie qui sera abandonnée au profit de la réparation (*repair*). Le *point d'interruption* (PI) établit la frontière finale du *reparandum* et marque une rupture dans la fluidité du discours. L'*interregnum* (IM) désigne la région comprise entre la frontière finale du *reparandum* et la frontière initiale du *repair*. Enfin, le *repair* (RR) représente la partie corrigée du *reparandum* et marque le retour à la « fluence » du discours. L'exemple suivant illustre ce cadre d'analyse :

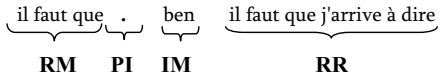


Figure 1 : Structure de la disfluence (Shriberg, 1994)

Cependant, ce modèle révèle un certain nombre de limites. Par exemple, la non-récursivité de ce modèle (*i.e* l'impossibilité d'avoir un schéma **RM/PI/IM/RR** à l'intérieur d'un premier schéma **RM/PI/IM/RR**) empêche de rendre compte de certaines configurations syntaxiques telle que l'imbrication de disfluences.

Il est très fréquent d'observer des imbrications de disfluences : une disfluence s'insère dans une autre avant que la première soit terminée créant ainsi une interdépendance entre les segments disfluents. Il s'agit en fait de plusieurs éléments sur l'axe paradigmatique qui se succèdent et qui se trouvent ainsi sous la dépendance les uns des autres. L'imbrication s'effectue au niveau de la syntaxe où l'on observe les unités syntaxiques fondées sur l'organisation des catégories grammaticales et de leur rection.

[on: on parlait souvent **[** du: du fameux euh coq **[** au: au Chambertin ↗ **]]]**

A l'inverse du modèle de Shriberg (1994) offrant une vision strictement linéaire de l'organisation des productions disfluentes, le modèle de la mise en grille proposé par (Blanche-Benveniste, 1987) permet de visualiser les configurations du discours grâce à une représentation qui suit deux axes : l'axe syntagmatique (horizontal) et l'axe paradigmatique (vertical). Les phénomènes de production sont ramenés à des « piétinements » sur une même place syntaxique.

on: on parlait souvent		
	du: du fameux euh coq	
		au: au Chambertin ↗

Figure 2 : Mise en grille de disfluence imbriquée

Dans cet exemple, les éléments *du fameux coq* et *au Chambertin* comportent des piétinements syntaxiques et sont rattachés en cascade au verbe recteur *parler*. Dans nos corpus, nous avons pu relever plusieurs imbrications plus ou moins compliquées. L'imbrication de trois disfluences successives représente le cas le plus complexe. L'intérêt réside ici en une représentation de l'architecture syntaxique des énoncés en suivant un cadre d'analyse unifié. La mise en grille complète la transcription du discours en la rendant à la fois plus lisible et plus compréhensible. Elle permet de traiter les disfluences avec une certaine neutralité (on ne « gomme » pas la disfluence). Grâce à cette représentation tous les essais de lexique sont conservés même s'ils ne font pas avancer le discours.

4 Schéma d'annotation

L'annotation des disfluences se révèle difficilement automatisable dans la mesure où l'étude de ces phénomènes réclame un jugement éminemment interprétatif de la part de l'annotateur. Pour cette raison, l'annotation se veut entièrement manuelle et s'effectue à l'aide du logiciel de transcription assistée par ordinateur *Transcriber* qui permet de lier les deux types de ressources nécessaires à l'exploitation des corpus oraux : le fichier son et la transcription.

4.1 Principes

Pour réaliser ce travail, il a fallu ensuite trouver une méthode suffisamment « robuste » pour délimiter et coder les disfluences de manière homogène, en limitant les ambiguïtés relatives aux choix d'annotation. Nous nous sommes inspirés du formalisme XML pour la création des balises délimitant les segments disfluents. Le choix d'un codage par balises présente plusieurs avantages : il facilite la hiérarchisation des informations et va dans le sens des normes actuelles qui privilégient ce type de codage dans les projets de normalisation et d'exploitation des corpus oraux (cf. Krul, 2002). De plus, ce type de codage représente un format d'échange standard « universel » et peut être ainsi intégré à des corpus existants utilisant déjà la norme XML.

Dans un premier temps, nous avons donc créé une balise encadrante **<dis>…</dis>** faisant office de délimiteur dans notre schéma d'annotation. Les « piétinements » syntaxiques propres au modèle de (Blanche-Benveniste, 1987) sont représentés par la balise « **<start/>** » placé devant chaque segment.

Exemple :

Segment disfluent initial :

en hiver au Portugal il p- il p- il y a des moments de pluie assez importants des fois ↗

Segment disfluent annoté (1ère passe) :

en hiver au Portugal
<dis>
<start/> il p-
<start/> il p-
<start/> il y a des moments de pluie assez importants des fois ↗
</dis>

Figure 3 : Première phase d'annotation

Dans un second temps, nous avons attribué à chaque segment disfluent, une étiquette pour qualifier le type de disfluence. Le fait d'affecter un type pour chaque disfluence a pour but de faciliter l'extraction d'informations ponctuelles telle que la répartition des types de disfluences dans le corpus. Là encore, nous nous sommes inspirée du formalisme XML pour décrire les propriétés des segments disfluents. Chaque disfluence possède un attribut « type » qui est défini à l'intérieur de la balise et d'une valeur associée. Nous définissons quatre valeurs possibles pour le type de disfluence : "rep" pour les répétitions, "ac" pour les autocorrections, "am" pour les amorces et "dc" pour les disfluences combinées.

<dis type=**"rep"**>
<start/> il y a:
<start/> il y a une re*mise en question
</dis>

Figure 4 : Deuxième phase d'annotation

Nous avons également annoté les constructions inachevées au moyen d'une marque ponctuelle « **<in/>** ». L'utilisation d'une marque ponctuelle est un choix plus judicieux pour annoter ce type de disfluence car il est très délicat de circonscrire l'inachèvement.

là c'était très bien hein j'étais au bord de la mer c'était super ça été un moment de: **<in/>**

Nous avons également ajouté les informations concernant les marqueurs discursifs (*bon*, *ben*, *voilà*, *donc*, etc. (Chanet, 2004)) sous la forme d'une balise encadrante **<md>…</md>**. Les marqueurs discursifs ne sont pas à proprement parler des disfluences (bien qu'étant étroitement liés à celles-ci), mais leur fréquence élevée dans notre corpus nous oblige à tenir compte de ces unités en les incluant dans le schéma d'annotation. D'un point de vue syntaxique, les marqueurs peuvent être définis comme des mots qui, dans le discours, n'entrent dans aucune construction syntaxique, tout en étant attachés prosodiquement au syntagme dans lequel ils prennent place.

<md>bon**</md>** là c'était très bien **<md>**hein**</md>** j'étais au bord de la mer c'était super

4.2 Problèmes rencontrés

Même en suivant un modèle théorique préalablement défini, l'annotation se heurte à des cas problématiques. La principale difficulté lorsque l'on annote les disfluences réside dans le fait qu'il n'est pas toujours évident de circonscrire le segment disfluent. En effet, l'absence de ponctuation dans les transcriptions peut poser problème pour délimiter le début et la fin de la disfluence. Un autre problème récurrent est celui des disfluences imbriquées où la difficulté principale réside dans l'application d'un balisage correct des disfluences les unes à l'intérieur des autres.

4.2.1 Délimitation du segment disfluent

Comme nous avons pu le voir dans les exemples précédents, la frontière gauche de la disfluence débute au premier « piétinement » syntaxique, ce qui ne pose pas de difficultés puisque l'on applique la même stratégie d'annotation pour chaque élément disfluent. En revanche, la borne à droite est beaucoup plus difficile à identifier. En l'absence de toute ponctuation, il nous semble que le rôle des indices prosodiques (mouvements intonatifs majeurs) et des autres marques comme les pauses et les allongements ainsi que les marqueurs discursifs est essentiel pour déterminer le début et la fin du segment disfluent. Nous nous sommes donc appuyés sur un certain nombre d'indices pour délimiter la frontière droite de la disfluence. Nous nous sommes basée en premier lieu, sur les éléments prosodiques présents dans le *CRFP*. Plusieurs auteurs ont fait mention du lien entre l'intonation et l'organisation syntaxique de l'énoncé :

> *« La présence de tons finals dominants, avec effet de regroupement, aux frontières syntaxiques majeures, indique une correspondance entre structure syntaxique et structure intonative. »* (Blanche-Benveniste, 1990 : 173)

La figure ci-après montre l'utilisation des marques intonatives comme délimitation de frontière droite de la disfluence.

```
<dis type="ac">
        <start/> mon
        <start/> un vieux collègue de sciences naturelles m'avait dit sur*tout ↗ + pas* d'histoire avec
        les filles ↘
</dis>

on a
<dis type="rep">
        <start/> des:
        <start/> des bons clients ↗
</dis>
```

Figure 5 : Délimitation de la frontière droite à partir de la prosodie

Cependant, les corpus oraux existants ne bénéficient pas systématiquement d'une annotation prosodique. Si l'on souhaite par exemple élargir notre méthodologie à d'autres corpus, il est donc nécessaire, à notre sens, de nous appuyer en parallèle sur d'autres indices plus largement codés dans les corpus oraux à savoir les pauses (silencieuses et remplies), les allongements et les marqueurs discursifs.

4.2.2 Disfluences imbriquées

Il est très fréquent d'observer des imbrications de disfluences : une disfluence s'insère dans une autre avant que la première soit terminée créant ainsi une interdépendance entre les segments disfluents. Il s'agit en fait de plusieurs éléments sur l'axe paradigmatique qui se succèdent et qui se trouvent ainsi sous la dépendance les uns des autres. L'imbrication s'effectue au niveau de la syntaxe où l'on observe les unités syntaxiques fondées sur l'organisation des catégories grammaticales et de leur rection.

```
<dis>
<start/> on:
<start/> on parlait souvent
                        <dis>
                        <start/> du:
                        <start/> du fameux euh coq
                                                <dis>
                                                <start/> au:
                                                <start/> au Chambertin ↗
                                                </dis>
                        </dis>
</dis>
```

Figure 6 : Imbrication de disfluences

Dans cet exemple, les éléments *du fameux coq* et *au Chambertin* comportent des piétinements syntaxiques et sont rattachés en cascade au verbe recteur *parler*. Dans nos corpus, nous avons pu relever plusieurs imbrications plus ou moins compliquées. L'imbrication de trois disfluences successives représente le cas le plus complexe.

4.2.3 Raccordement syntaxique impossible

Notre méthode d'annotation permet de rendre compte de la plupart des configurations de disfluences. Cependant, quelques énoncés disfluents ne peuvent pas être annotés en utilisant

cette méthode. En effet, les relations entre les éléments peuvent être à une distance tout à fait notable et le raccordement syntaxique s'effectue trop loin dans l'énoncé pour être annoté et mis en grille. C'est le cas notamment de l'exemple suivant pour lequel il n'a pas été possible de mettre en place notre schéma d'annotation :

ben parce que le: Charlemagne euh + paraît-il ↗ **ne buvait que des rouges** ↗ + parce qu'
<dis type="dc">
<start/> il voulait pas ↗
<start/> il
<start/> il se tâchait ↗ +
<dis type="dc">
<start/> sa:
<start/> sa: + </in>
</dis>
</dis> il **ne buvait que* des Blancs** pardon ↗

Figure 7 : Autocorrection non annotée

La correction *ne buvait que des Blancs* remplace *ne buvait que des rouges* initialement produite par le locuteur. Cependant, cet énoncé se situe à une trop grande distance pour être annoté en tant qu'autocorrection.

5 Aspects quantitatifs

Après avoir annoté les corpus, nous avons réalisé une étude quantitative dont l'objectif est d'illustrer la typologie des phénomènes de l'oral. L'approche quantitative permet d'accéder plus facilement à la description des phénomènes qui présentent de l'intérêt et dont il aurait été difficile de cerner les contours *a priori*. A l'aide de scripts permettant d'automatiser les décomptes (scripts en *Perl* et *Bash*), nous avons quantifié les types de disfluences, les constructions inachevées ainsi que les marqueurs discursifs.

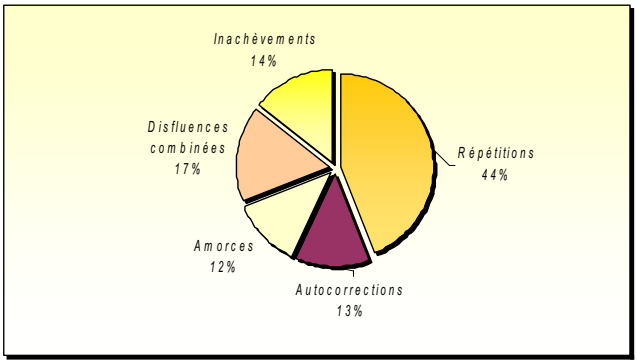


Figure 8 : Répartition des types de disfluences

Nous avons relevé 293 disfluences : les répétitions constituent le type le plus largement représenté (44%). Les autres types sont repartis de manière plus homogène : leur pourcentage varie entre 17% pour les disfluences combinées et 12% pour les amorces. Les comptages

permettent également d'effectuer quelques constats sur la fréquence des marqueurs discursifs, et permettent de dégager des hypothèses sur le fonctionnement de ces unités.

TÊTE DE LISTE : Fréquence des marqueurs discursifs			
Rang	Forme	Fréquence	Fréquence Relative
1	donc	67	18,61%
2	hein	55	15,28%
3	bon	36	10%
4	quand même	15	4,17%
5	enfin	15	4,17%
6	ben	13	3,61%
7	alors	13	3,61%
8	là	9	2,50%
9	quoi	8	2,22%
10	mais	8	2,22%

Tableau 1: Fréquence des marqueurs discursifs (tête de liste)

6 Conclusion et perspectives

L'étude de l'oral est aujourd'hui un thème de recherche très riche, même s'il reste encore de nombreux progrès à faire pour permettre d'automatiser complètement son traitement. L'une des premières phases dans l'optique du développement d'applications en TAL dans ce domaine, peut passer, notamment, par la constitution et l'annotation de corpus oraux.

Cet article rend compte du travail d'annotation effectués de ces phénomènes à partir de corpus oral (*Corpus de Référence du français parlé*) en suivant un modèle d'analyse précis (la mise en grille) et en utilisant une norme générique d'annotation de textes (XML). De plus, l'étude quantitative et qualitative menées conjointement dans notre travail nous ont permis d'avoir une connaissance plus précise des caractéristiques des phénomènes de l'oral même s'il reste encore beaucoup de cas à étudier, et ce, sur de plus grands volumes de données. Nous avons pu, grâce aux observations sur corpus, dégager un certain nombre de régularités qui nous renseignent sur le fonctionnement des disfluences (qui peut être enrichi par l'observation des différents patrons syntaxiques des segments disfluents (Piu, 2006)) et peuvent servir de base pour l'amélioration des outils de traitement automatique (par exemple l'analyse syntaxique automatique) qui butent encore sur les données orales.

Références

BLANCHE-BENVENISTE, C., & JEANJEAN, C. (1987). *Le français parlé – Édition et transcription.* Paris : Didier-Érudition.

BLANCHE-BENVENISTE, C. (1990). *Le français parlé – Études grammaticales.* Paris : CNRS.

CANDÉA, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané.* Thèse de doctorat. Université Paris III.

CHANET, C. (2004). *Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie.* Recherches sur le français parlé, 18, 83-105.

ÉQUIPE DELIC. (2004). Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé* 18, 11-43.

HENRY, S. (2002). Étude des répétitions en français parlé spontané pour les technologies de la parole. Actes de la *6ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, 467-476. Nancy (France).

KRUL, A. (2002). *Annotation structurelle de corpus oraux avec XML*. Mémoire de Maîtrise. Université Paris III Sorbonne Nouvelle.

KURDI, M. Z. (2003). *Contribution à l'analyse du langage oral spontané.* Thèse de doctorat. Université de Grenoble I.

PALLAUD, B. (2002). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le Français parlé*, 17, 79-102.

PIU, M. (2006). *Annotation des disfluences dans les corpus oraux*. Mémoire de Master. Université de Provence, Aix-en-Provence.

VÉRONIS, J. (1998). *Annotation automatique de corpus : état de la technique*. Colloque International « Questions de méthode dans la linguistique de corpus ». Perpignan (France).

Architecture modulaire portable pour la génération du langage naturel en dialogue homme-machine

Vladimir POPESCU[1,2]
[1] Laboratoire d'Informatique de Grenoble, France
[2] Université « Politehnica » de Bucarest, Roumanie
Vladimir.Popescu@imag.fr

Résumé. La génération du langage naturel pour le dialogue oral homme-machine pose des contraintes spécifiques, telles que la spontanéité et le caractère fragmenté des énoncés, les types des locuteurs ou les contraintes de temps de réponse de la part du système. Dans ce contexte, le problème d'une architecture rigoureusement spécifiée se pose, autant au niveau des étapes de traitement et des modules impliqués, qu'au niveau des interfaces entre ces modules. Afin de permettre une liberté quasi-totale à l'égard des démarches théoriques, une telle architecture doit être à la fois modulaire (c'est-à-dire, permettre l'indépendance des niveaux de traitement les uns des autres) et portable (c'est-à-dire, permettre l'interopérabilité avec des modules conçus selon des architectures standard en génération du langage naturel, telles que le modèle RAGS - « Reference Architecture for Generation Systems »). Ainsi, dans cet article on présente de manière concise l'architecture proposée, la comparant ensuite au modèle RAGS, pour argumenter les choix opérés en conception. Dans un second temps, la portabilité de l'architecture sera décrite à travers un exemple étendu, dont la généralité réside dans l'obtention d'un ensemble de règles permettant de plonger automatiquement les représentations des informations de notre architecture vers le format du modèle RAGS et inversement. Finalement, un ensemble de conclusions et perspectives clôturera l'article.

Abstract. Natural language generation for human-computer dialogue imposes specific constraints, such as the spontaneous and fragmented character of the utterances, speaker types or constraints related to the system's time of response. In this context, the issue of a thouroughly specified architecture stems naturally, with respect to the processing stages in the modules involved and to the interfaces between these modules as well. In order to allow for a quasi-total freedom concerning the theoretical principles driving the processing stages, such an architecture must be modular (i.e., allowing the independence of the modules of each other) and portable (i.e., allowing a certain interoperability between modules designed following this architecture and existing modules, designed following standard, reference architectures, such as the RAGS model). Thus, in this paper firstly the proposed architecture will be presented in a concise manner, comparing it then to the RAGS model and arguing for the design choices being made. Afterwards, the portability of the architecture will be described, via an extended example whose general character resides in the fact that a set of rules are obtained, that allow automatic translations between representation formats in our architecture and in the RAGS model, in both ways. Finally, a set of conclusions and pointers to further work end up the paper.

Mots-clés : génération, dialogue, architecture modulaire, portabilité, XML.

Keywords: generation, dialogue, modular architecture, portability, XML.

1 Introduction

Nos recherches se situent dans le cadre de la génération du langage naturel pour le dialogue oral homme-machine et concernent le développement d'un module de génération du langage naturel pour donner un caractère aussi « naturel » et expressif que possible aux réponses langagières du système face aux requêtes des usagers. Ce travail poursuit ainsi des recherches commencées depuis plusieurs années (Imberdis & Caelen, 1997). Le problème n'est pas simple, car la plupart des générateurs textuels existants sont conçus pour des situations de monologue, et il s'y ajoutent le caractère spontané et fragmenté du dialogue, auquel des contraintes de pertinence, expressivité et temps de réponse se conjuguent (McTear, 2002).

Ainsi, nous considérons (Popescu *et al.*, 2007) que la réponse du système se situe à cinq niveaux auxquels le langage naturel peut être « produit » par un système de dialogue : (i) le niveau **logique** instantié dans un contrôleur de dialogue et ne faisant pas partie du générateur, mais fournissant l'intention communicationnelle à mettre sous forme linguistique, (ii) le niveau **pragmatique** gérant les aspects liés à l'expressivité des énoncés et à leur pertinence par rapport au contexte dialogique, (iii) le niveau **linguistique** produisant le texte pour l'intention communicationnelle, (iv) le niveau **expressif** calculant la forme finale de l'énoncé et la prosodie, et (v) le niveau **acoustique** réalisant la synthèse de la parole proprement dite.

L'architecture conçue pour la génération dans le cadre du dialogue homme-machine part d'un ensemble de principes :

1. prendre en compte les aspects pragmatiques (rhétoriques - structuration discursive et expressifs - valences émotionnelles) et la gestion des tours de parole (dialogue oral spontané), ainsi que les particularités des locuteurs ;
2. considérer surtout les performances des agents du dialogue (en dépit des normes de compétence qui varient d'un contexte (social, situationnel, etc.) à l'autre et ne sont donc pas génériques) ;
3. s'appuyer sur un corpus de dialogues réels entre humains et entre homme et machine (en dépit des prescriptions grammaticales fixées a priori) pour contrôler les aspects linguistiques en génération ;
4. rendre les traitements appropriés à un fonctionnement en temps réel (et donc éviter des processus d'inférence relativement coûteux) ;
5. diminuer autant que possible la dépendance à la tâche et à la langue, en permettant des paramétrages aisés ;
6. rendre les niveaux de traitement de la parole aussi indépendants que possible des traitements purement textuels.

Le premier point des désidérata ci-dessus est réalisé dans le générateur pragmatique, où les aspects rhétoriques-discursifs sont gérés par l'utilisation adaptée de la théorie SDRT (« Segmented Discourse Representation Theory ») (Asher & Lascarides, 2003), tandis que les aspects expressifs au niveau de chaque énoncé sont gérés par le contrôle du « degré de puissance de la force illocutoire ». De plus, on prévoit la prise en compte, dans le générateur pragmatique, d'un doublet d'attributs caractérisant chaque partenaire du dialogue (en termes de niveau de familiarité par rapport au dialogue courant et de relation sociale de l'un avec l'autre).

Le cinquième point est réalisé par le fait que, autant au niveau pragmatique que linguistique, les méthodes utilisées sont indépendantes de la tâche et de la langue (supposant en fin de compte

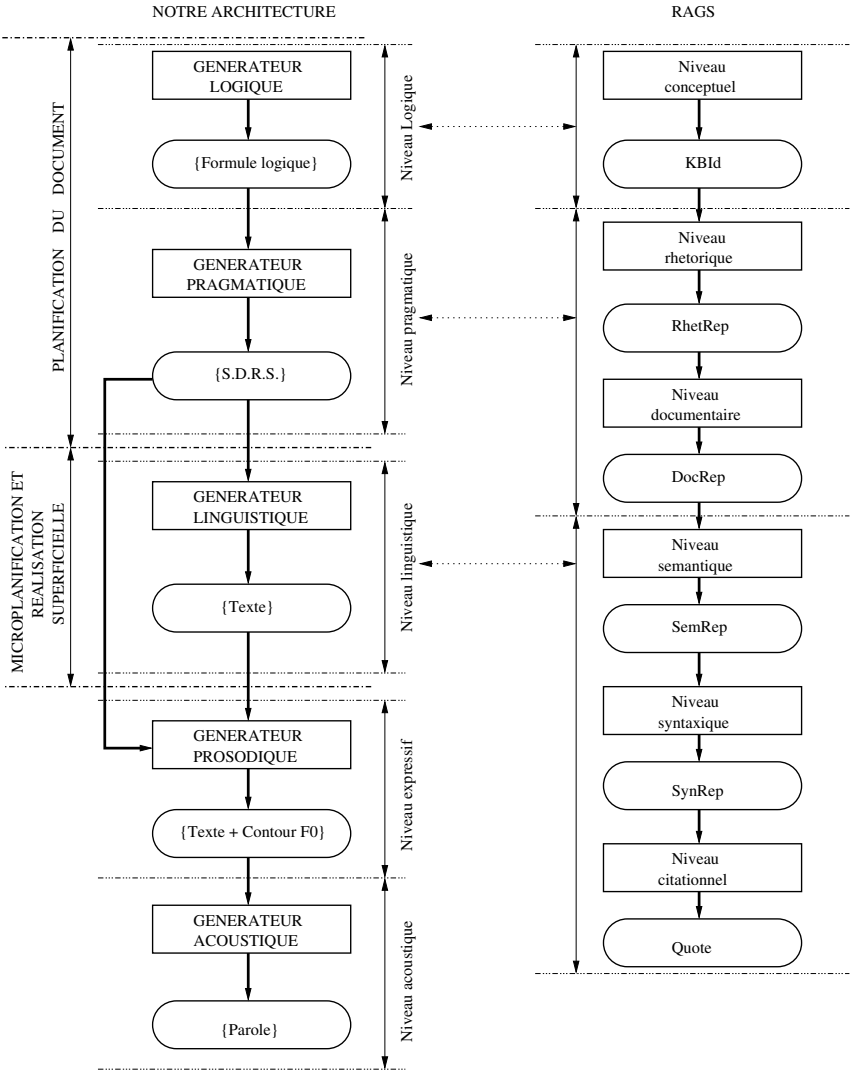


FIG. 1 – Architecture modulaire de génération et correspondances avec le modèle RAGS

des traitements du type appariement des graphes) et paramétrables dans une langue et pour une tâche données.

Le sixième point est réalisé par l'utilisation d'un format standard pour la représentation des connaissances (XML) et d'une théorie de représentation pragmatique-discursive dont le formalisme « interne » n'est pas pris en compte (la SDRT - « Segmented Discourse Representation

Theory », dont on ne prend en compte que les relations rhétoriques et leurs sémantiques *informelles*).

L'idée de concevoir une architecture pour la génération du langage naturel a été déjà énoncée dans plusieurs études, comme celle de Reiter et Dale (Reiter & Dale, 2000) et l'architecture RAGS (Mellish *et al.*, 2006). Mais aucun de ces deux modèles ne traite le dialogue homme-machine. Il existe peu de travaux pour définir une architecture adaptée pour la génération en dialogue (Imberdis & Caelen, 1997), (McTear, 2002).

D'autres travaux ont abordé la génération en dialogue de diverses façons, dont les plus notables sont, à notre sens, ceux d'Amanda Stent (Stent, 2001), de Mariet Theune (Theune, 2000) et de Matthew Stone (Stone, 1998), puisqu'ils sont explicitement concernés par le dialogue oral homme-machine, aboutissant en même temps à des systèmes fonctionnels. Ces démarches ne sont pas génériques, sont sans référence à un « standard » en génération (tels que l'architecture de Reiter et Dale ou le modèle RAGS). Cela implique le manque de portabilité et de réutilisation de ces systèmes. Notre architecture proposée ici est fortement compatible à la fois avec le propos de Reiter et Dale et avec le modèle RAGS.

Pour l'utilisation de la SDRT en génération en situations de monologues, les travaux de Laurence Danlos et son équipe peuvent être citées (Danlos *et al.*, 2001) ; pour des extensions de la SDRT pour l'interprétation des dialogues, les travaux de Laurent Prévot (Maudet *et al.*, 2004) et son équipe sont intéressants, mais ne concernent pas la génération. Dans ce contexte, les travaux assumés par notre projet essaient de renforcer les recherches en génération pour le dialogue homme-machine finalisé.

2 Architecture modulaire compatible avec les « standards » en génération

Le schéma global de l'architecture est illustré dans la figure 1, où on met en évidence la relation entre les niveaux de traitement, les modules (générateurs) et les connaissances échangées entre les *modules* et implicitement entre les niveaux.

Sur cette figure les modules de traitement sont représentés par des rectangles et les structures de données par des ellipses. A droite, les flèches doubles montrent les niveaux de traitement, tandis qu'à gauche, les flèches situent l'architecture proposée ici par rapport à l'architecture de référence de Reiter et Dale (Reiter & Dale, 2000), qui représente une première spécification, plutôt théorique, d'une architecture pour la génération du langage naturel.

Les interfaces entre les modules de génération sont spécifiées en XML et les choix de conception pour notre architecture ont été décrites de manière étendue dans (Popescu *et al.*, 2007). Le modèle RAGS est décrit dans (Mellish *et al.*, 2006). Les entités dénommées « primitives abstraites » à un niveau de traitement dans la description du modèle RAGS ci-dessus sont en fait des pointeurs renvoyant vers les représentations les plus complexes au niveau inférieur ; par exemple, la primitive RhetLeaf au niveau rhétorique renvoie vers une structure sémantique, SemRep.

Le modèle RAGS spécifie une architecture « pipeline », permettant des raccourcis entre les niveaux de représentation, via les flèches *non-locales* (Mellish *et al.*, 2006). Donc, la double incidence du niveau expressif à d'autres niveaux de traitement est compatible en théorie et en

pratique avec RAGS.

Les correspondances entre RAGS et notre architecture sont montrées dans la figure 1, où les connaissances transférées dans le modèle RAGS sont des primitives (dans le cas de KBId et Quote) où des représentations dérivées (pour le reste - cf. la discussion ci-dessous).

Notre architecture

- **Niveau Logique** - Ce niveau de traitement ne fait pas partie à vrai dire de la composante de génération d'un système de dialogue, mais on le précise ici en raison de compatibilité avec les architectures existantes dans le domaine de la génération ; il fournit une forme logique correspondante à l'intention communicationnelle à mettre sous forme linguistique, prenant en entrée une formule logique correspondante à l'intention communicationnelle d'un énoncé provenu de l'usager ;
- **Niveau pragmatique** - Ce niveau de traitement est le premier à faire vraiment partie du système de génération ; il utilise la SDRT et les actes de langage pour fournir une représentation discursive du dialogue courant, prenant en entrée la formule logique pour l'énoncé à engendrer ;
- **Niveau linguistique** - Ce niveau de traitement opère la mise en forme linguistique du message à générer par le système et prend en entrée la structure discursive pour le dialogue courant, fournissant le texte correspondant à l'énoncé matérialisant la réponse du système ;
- **Niveau expressif** - Ce niveau de traitement ajoute un degré d'expressivité au texte engendré au niveau linguistique, prenant en entrée la structure discursive pour fournir en sortie les paramètres prosodiques appropriés associés au texte à générer ; l'expressivité est considérée seulement à l'égard de la qualité du signal vocal synthétisable à partir du texte ;
- **Niveau acoustique** - Ce niveau réalise la synthèse vocale proprement dite, en prenant à l'entrée un texte annoté pour une gestion fine de la prosodie pour fournir en sortie la parole synthétisée ;

Le modèle RAGS

- **Niveau conceptuel** - Ce niveau de traitement fournit des représentations non-linguistiques des informations à communiquer (par exemple des formules logiques) ; ces représentations conceptuelles sont manipulées via des pointeurs vers des entités primitives dans une base de connaissances et appelées KBId (« Knowledge Base Identifier ») ;
- **Niveau rhétorique** - Ce niveau de traitement définit des relations discursives entre les énoncés enchaînés dans un discours ; la manière dont les relations sont réalisées est spécifiée via les primitives abstraites RhetRel (pour la relation) et RhetLeaf (pour les arguments de la relation) à partir desquelles les structures composites RhetRepSeq et RhetRep sont construites ;
- **Niveau documentaire** - Ce niveau de traitement concerne la disposition physique des énoncés (voire typage du texte) à produire et les structures de données fournies sont DocAttr, DocLeaf, DocRepSeq et DocRep, construites à partir des primitives abstraites DocFeat, DocAtom et DocLeaf ;
- **Niveau sémantique** - Ce niveau de traitement peut être combiné avec le niveau rhétorique, ainsi qu'au niveau sémantique on code à la fois les propositions atomiques (cf. celles trouvées dans les structures rhétoriques) et les organisations structurées de propositions, pour l'entrée aux générateurs dits « à entrée sémantique » ; les structures de données fournies par ce niveau, SemRep, ScopedSemRep, SemType, SemAttr,Scoping et ScopeConstr sont construites à partir des primitives abstraites DR, SemConstant, SemPred et ScopeRel ;
- **Niveau syntaxique** - Ce niveau fournit une représentation syntaxique abstraite qui ne spécifie pas l'ordre des mots à l'intérieur des énoncés ; par rapport à la représentation sémantique, la structure syntaxique comprend les fonctions grammaticales des mots et est représentée par les types SynRep, FVM (de « Feature-value matrix », matrice attribut-valeur), SynArg, et Adj, construites à partir des primitives nil, SynFun, SynFeat et SynAtom ;
- **Niveau citationnel** - Ce niveau concerne surtout les fragments fixés de texte (les structures figées) incluses, sans modification, dans la sortie du système de génération ; pour la représentation des données à ce niveau seule une primitive abstraite existe, Quote ;

L'architecture proposée ici et son parallèle, l'architecture RAGS sont comparés ci-dessous.

Ainsi, le niveau logique de notre architecture correspond au niveau conceptuel du modèle RAGS seulement du point de vue de la génération, car le générateur logique (en fait, le contrôleur de dialogue) a un rôle plus étendu que le niveau conceptuel de RAGS. C'est pour cela que le niveau logique ne fait pas à vrai dire partie de notre architecture, mais représente seulement l'interface du système de génération au contexte dialogique.

Le niveau pragmatique de notre architecture correspond au niveau rhétorique de RAGS et au niveau documentaire de ce dernier, car notre générateur pragmatique assure la situation et la cohérence rhétorique de l'acte de langage à engendrer, en fournissant en même temps une représentation qui constitue une structuration discursive.

Le niveau linguistique de notre architecture correspond à trois niveaux de RAGS : les niveaux

sémantique, syntaxique et citationnel.

Quant aux niveaux expressif et acoustique de notre architecture, ceux-ci n'ont pas de correspondants en RAGS car le côté prosodie et synthèse de parole représente un aspect de la génération dont RAGS ne rend pas compte. RAGS est concerné seulement avec l'obtention du texte comme résultat final du processus de génération.

En conclusion, on observe que l'architecture, bien qu'elle s'appuie sur d'autres architectures de référence, soit plus théoriques comme celle de Reiter et Dale, soit plus orientées logiciel, comme RAGS, reste particulière dans les détails fins.

3 Architecture portable

Nous discutons maintenant de l'isomorphisme entre les représentations de RAGS et les représentations de notre architecture, en nous appuyant sur un exemple de représentations dans le modèle RAGS et dans notre architecture.

Nous illustrons cela autour d'un exemple pour la représentation rhétorique correspondante à un dialogue court entre deux locuteurs, ayant comme objet l'établissement d'un rendez-vous :

$U : \pi_1$: Quand es-tu disponible aujourd'hui ?

$M : \pi_2$: Je suis disponible aujourd'hui dès 16h !

Ainsi, en employant une formalisation rhétorique à la SDRT (Asher & Lascarides, 2003), on suppose que les générateurs doivent représenter la structuration rhétorique du dialogue indiqué ci-dessus.

Pour ce texte, la représentation rhétorique RhetRep que le niveau rhétorique de RAGS construit dans son langage interne est montrée dans la figure 2, à côté de la représentation issue de notre architecture, dans le cadre de la SDRT. Les conventions de représentation pour le RAGS sont les suivantes (Mellish *et al.*, 2006) : (i) les noeuds sont des *objets* qui représentent des structures et des sous-structures (types primitifs) ; à chaque objet on associe un type RAGS (de ceux définis dans le modèle) ; de ces types, seulement les types primitifs spécifient des informations se reportant à la théorie choisie pour *concrétiser* l'architecture abstraite de RAGS, et (ii) les *flèches* entre les objets représentent les relations entre ces derniers ; pour chaque flèche il existe une et seulement une source et une cible ; en même temps, chaque flèche est désignée par une étiquette unique. De plus, les flèches sont typées, dans le sens que pour chaque étiquette il y a un unique *type* pour la source et un type toujours unique pour la cible ; ces types sont précisés lors de la définition de chaque flèche, donc, immuables. En revanche, pour la représentation de notre architecture, la notation « > » sépare deux éléments en structure arborescente, tandis que « : : » sépare un élément (à gauche) d'un attribut des siens propres.

On peut voir le texte ci-dessus relié par la relation discursive SDRT QAP (« Question - Answer Pair »), dont la représentation RAGS est illustrée dans la figure 2.

La représentation XML correspondant à la représentation de la figure 2, pour cet exemple, est donnée dans la figure 3.

La figure 4 montre un ensemble de règles permettant le passage de la représentation rhétorique de RAGS à la représentation utilisée en sortie du module pragmatique de notre architecture ; on utilise les mêmes notations concernant « > » et « : : ».

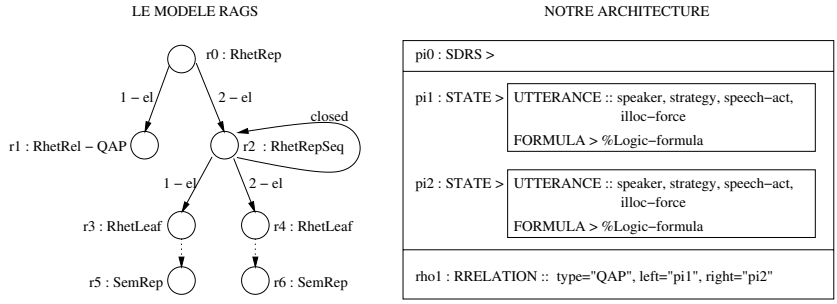


FIG. 2 – Exemple de représentation rhétorique dans RAGS et dans notre architecture pour une instance de dialogue simple

Cet ensemble de règles montre que pour 23 des 34 règles il existe une correspondance entre les éléments et attributs dans la représentation RAGS et dans notre architecture, donc il y a une superposition de 2/3 entre les représentations dans les deux architectures au sens où la représentation dans notre architecture couvre 2/3 de la représentation équivalente en RAGS. Le reste comprend surtout des informations d'identification interne des structures de données, ce qui n'est pas nécessaire dans notre architecture, car elle est conçue pour le dialogue homme-machine où le système doit générer un tour de parole à un instant donné et non pas élaborer des discours monologiques étendus, où un plus fort repérage des structures linguistiques partielles est nécessaire dans le temps.

Procédant à l'envers et construisant un ensemble de règles permettant de « traduire » la spécification XML des informations rhétoriques échangées entre les modules de génération pragmatique et linguistique en représentation selon le modèle RAGS on obtient une couverture de 5/6 selon un ensemble de règles semblables à ceux juste présentées ci-dessus. Le reste de 1/6 relève du manque de « transport » vers la représentation RAGS des marqueurs spécifiques au dialogue (l'identité des locuteurs et les marqueurs pragmatiques dialogiques - stratégie, acte de langage et force illocutoire). Ces informations peuvent cependant être récupérées au niveau des identifiants en RAGS, qu'une structure parallèle réalisant les correspondances appropriées peut accompagner, en tant que connaissances *statiques* sous forme de tableau de correspondances.

4 Conclusion

Dans cet article nous avons présenté une vue d'ensemble d'architecture **générique** (dans le sens que les modules de traitement sont indépendants des *interfaces* entre eux), **modulaire** (dans le sens que les traitements dans un module sont indépendants des traitements dans les autres modules) et **portable** (dans le sens que même au niveau des interfaces les modules sont indépendants pourvu qu'on plonge les représentations des connaissances à chaque niveau de traitement dans une architecture standard en génération du langage naturel, telle que le modèle RAGS).

Plus précisément, un exemple étendu où la représentation de l'interface entre les modules pragmatique et linguistique dans notre architecture, comportant des informations rhétoriques échan-

LE MODELE RAGS

```
<RAGS>
 <RhetRep ident="id1" type="tuple"
          length="2">
  <RhetRel name="QAP"/>
  <RhetRepSeq ident="i2" type="sequence">
   <RhetLeaf ident="ID3"/>
   <RhetLeaf ident="ID4"/>
  </RhetRepSeq>
 </RhetRep>
 <arrow name="refers_to" source="ID3"
        target="ID5"/>
 <arrow name="refers_to" source="ID4"
        target="ID6"/>
 <SemRep ident="ID5" type="tuple"
         length="3">
  <DR name="r5"/>
  <SemType type="set">
   <SemPred name="available"/>
  </SemType>
  <SemAttr type="functional">
   <SemRoleRep>
    <SemRole name="actor"/>
    <SemConstant ident="C1"
                 name="patient"/>
   </SemRoleRep>
   <SemRoleRep>
    <SemRole name="actee">
     <SemRep ident="s2" type="tuple"
             length="3">
      <DR name="r7"/>
      <SemType type="set">
       <SemPred name="tu"/>
      </SemType>
      <SemAttr type="functional">
       <SemRoleRep>
        <SemRole name="person"/>
        <xref idref="C1"/>
       </SemRoleRep>
      </SemAttr>
     </SemRole>
    </SemRep>
   </SemRoleRep>
  </SemAttr>
 </SemRep>
 <SemRep ident="ID6" type="tuple"
         length="3">
  <DR name="r6"/>
  <SemType type="set">
   <SemPred name="16:00"/>
  </SemType>
  <SemAttr type="functional">
   <SemRoleRep>
    <SemRole name="subject"/>
    <xref idref="S2"/>
   </SemRoleRep>
  </SemAttr>
 </SemRep>
</RAGS>
```

NOTRE ARCHITECTURE

```
<SDRS label="pi0" nstates="2" nrels="1">
 <STATE label="pi1" type="UTT">
  <UTTERANCE speaker="A" strategy="K2"
             speech-act="FFS" illoc-force="36"/>
  <FORMULA>
   <quant name="exists" variable="X">
    <type>person</type>
    <quant name="exists" variable="Y">
     <type>time</type>
     <conn name="and">
      <pred name="equals">
       <term name="X"/> <term name="A"/>
      </pred>
      <pred name="equals">
       <term name="Y"/> <term name="?"/>
      </pred>
      <pred name="available">
       <term name="X"> <term name="Y">
      </pred>
     </conn>
    </quant>
   </quant>
  </FORMULA>
 </STATE>
 <STATE label="pi2" type="UTT">
  <UTTERANCE speaker="B" strategy="K2"
             speech-act="FS" illoc-force="54"/>
  <FORMULA>
   <quant name="exists" variable="X">
    <type>person</type>
    <quant name="exists" variable="Y">
     <type>time</type>
     <quant name="exists" variable="Z">
      <type>offset</type>
      <quant name="exists" variable="T">
       <type>direction</type>
       <conn name="and">
        <pred name="equals">
         <term name="X"/> <term name="A"/>
        </pred>
        <pred name="equals">
         <term name="Y"/> <term name="16:00+Z"/>
        </pred>
        <pred name="equals">
         <term name="Z"/> <term name="0"/>
        </pred>
        <pred name="equals">
         <term name="T"/> <term name="t+"/>
        </pred>
       </conn>
      </quant>
     </quant>
    </quant>
   </quant>
  </FORMULA>
 </STATE>
 <RRELATION label="rho1" type="QAP"  left="pi1" right="pi2"/>
</SDRS>
```

FIG. 3 – Représentations XML des informations rhétoriques, dans le modèle RAGS et dans notre architecture

1.	RAGS	→	∅
2.	RAGS > RhetRep	→	SDRS
3.	RAGS > RhetRep : : ident	→	SDRS : : label
4.	RAGS > RhetRep : : type	→	∅
5.	RAGS > RhetRep : : length	→	SDRS : : nstates
6.	RAGS > RhetRep > RhetRel	→	SDRS > RRELATION
7.	RAGS > RhetRep > RhetRel : : name	→	SDRS > RRELATION : : type
8.	RAGS > RhetRep > RhetRepSeq	→	SDRS > RRELATION
9.	RAGS > RhetRep > RhetRepSeq : : ident	→	SDRS > RRELATION : : label
10.	RAGS > RhetRep > RhetRepSeq : : type	→	∅
11.	RAGS > RhetRep > RhetRepSeq > RhetLeaf	→	SDRS > STATE
12.	RAGS > RhetRep > RhetRepSeq > RhetLeaf : : ident	→	SDRS > STATE : : label | SDRS > RRELATION : : left | SDRS > RRELATION : : right
13.	RAGS > arrow	→	∅
14.	RAGS > arrow : : name	→	∅
15.	RAGS > arrow : : source	→	∅
16.	RAGS > arrow : : destination	→	∅
17.	RAGS > SemRep	→	SDRS > STATE | SDRS > STATE > FORMULA
18.	RAGS > SemRep : : ident	→	SDRS > STATE : : label
19.	RAGS > SemRep : : type	→	SDRS > STATE : : type
20.	RAGS > SemRep : : length	→	∅
21.	RAGS > SemRep > DR	→	SDRS > STATE
22.	RAGS > SemRep > DR : : name	→	SDRS > STATE : : label
23.	RAGS > SemRep > SemType	→	SDRS > STATE | SDRS > STATE > FORMULA
24.	RAGS > SemRep > SemType : : type	→	SDRS > STATE : : type
25.	RAGS > SemRep > SemType > SemPred	→	SDRS > STATE > FORMULA > quant > conn > pred
26.	RAGS > SemRep > SemType > SemPred : : name	→	SDRS > STATE > FORMULA > quant > conn > pred : : name
27.	RAGS > SemRep > SemAttr	→	∅
28.	RAGS > SemRep > SemAttr : : type	→	∅
29.	RAGS > SemRep > SemAttr > SemRoleRep	→	SDRS > STATE > FORMULA | SDRS > STATE > FORMULA > quant | SDRS > STATE > FORMULA > quant > conn
30.	RAGS > SemRep > SemAttr > SemRoleRep > SemRole	→	SDRS > STATE > FORMULA > quant > conn > pred > term | SDRS > STATE > FORMULA > quant
31.	RAGS > SemRep > SemAttr > SemRoleRep > SemRole : : name	→	SDRS > STATE > FORMULA > quant > type
32.	RAGS > SemRep > SemAttr > SemRoleRep > Constant	→	∅
33.	RAGS > SemRep > SemAttr > SemRoleRep > Constant : : ident	→	SDRS > STATE > FORMULA > quant > variable
34.	RAGS > SemRep > SemAttr > SemRoleRep > Constant : : name	→	SDRS > STATE > FORMULA > quant > type

FIG. 4 – Règles de transformation du format XML RAGS vers celui de notre architecture

gées, a été plongée dans la représentation rhétorique, située au niveau conceptuel de traitement, dans le modèle RAGS. Lors de cet exercice, rendu possible via un ensemble de règles transformationnelles, on a constaté que la « traduction » dans les deux formats de représentation n'est pas totale, mais conserve les éléments pertinents par rapport à la nature de l'échange, dans les deux sens, de notre architecture vers RAGS et à l'inverse. Ce résultat nous amène a démontrer la portabilité annoncée : les modules dans notre architecture peuvent être combinés ou « entrelacés » avec des modules conçus suivant les spécifications du modèle RAGS ; par exemple, on pourrait construire un système de génération du langage naturel ayant les niveaux logique et pragmatique de notre architecture, ayant aussi un niveau linguistique éclaté en niveaux sémantique, syntaxique et citationnel d'un module de génération suivant les spécifications du modèle RAGS.

On a donc présenté une architecture qui étend un modèle de génération automatique des textes (RAGS) au dialogue oral homme-machine, en gardant, dans une mesure importante, la compatibilité entre les deux.

Nous envisageons de concevoir des algorithmes pour l'obtention automatique des règles de traduction entre notre architecture et le modèle RAGS ; ceci peut se réaliser en principe par l'appariement d'arbres XML pour les représentations dans les deux architectures ; la suite des

transformations d'un arbre vers l'autre constitue l'ensemble des règles. Evidemment, les démarches décrites dans cet article concernant l'interface entre les niveaux pragmatique et linguistique dans notre architecture et, respectivement, l'interface entre les niveaux rhétorique et sémantique dans le modèle RAGS doivent être étendues aux autres interfaces entre les modules des deux architectures.

Remerciement

L'auteur remercie vivement Jean Caelen, du Laboratoire d'Informatique de Grenoble, pour ses conseils attentifs et pour son apport aux travaux présentés dans cet article.

Références

ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge University Press.

DANLOS L., GAIFFE B. & ROUSSARIE L. (2001). Document structuring à la sdrt. In ACL, Ed., *Proceedings of the 8th European Workshop on Natural Language Generation EWNLG 2001*.

IMBERDIS L. & CAELEN J. (1997). Génération d'actes illocutoires pour le dialogue. In *Actes du Colloque Génération automatique du texte GAT'97*, Grenoble.

MAUDET N., MULLER P. & PRÉVOT L. (2004). Tableaux conversationnels en sdrt. In *Workshop SDRT, TALN 2004*.

MCTEAR M. F. (2002). Spoken language technology : Enabling the conversational user interface. In *ACM Computer Surveys 34 (1)* : ACM.

MELLISH C., SCOTT D., CAHILL L., PAIVA D., EVANS R. & REAPE M. (2006). A reference architecture for natural language generation systems. In *Journal of Natural Language Engineering 12 (1)*, p. 1–34 : Cambridge University Press.

POPESCU V., CAELEN J. & BURILEANU C. (2007). Generic architecture for natural language generation in spoken human-computer dialogue. In C. BURILEANU, Ed., *The 4th Conference on Speech Technology and Human-Computer Dialogue SpeD 2007* : Romanian Academy Publishing House.

REITER E. & DALE R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

STENT A. (2001). *Dialogue Systems as Conversational Partners : Applying Conversation Acts Theory to Natural Language Generation for Task-Oriented Mixed-Initiative Spoken Dialogue*. Ph D Thesis, University of Rochester.

STONE M. (1998). *Modality in Dialogue : Planning, Pragmatics and Computation*. Ph D Thesis, University of Pennsylvania.

THEUNE M. (2000). *From Data to Speech : Language Generation in Context*. Ph D Thesis, University of Eindhoven.

Résolution anaphorique intégrée à une analyse automatique de discours d'un corpus oral retranscrit

Alain RÉGNIER
LPL CNRS, Université de Provence,
29 Avenue R. Schuman 13621 Aix-en-Provence
alain.regnier@lpl.univ-aix.fr

Résumé. Nous présentons une résolution anaphorique intégrée à une analyse automatique de discours. Cette étude traite des anaphores pronominales et des anaphores zéro. Notre analyse est basée sur trois approches : une analyse basée sur les contraintes, une analyse fonctionnelle et une analyse dynamique. Pour évaluer la faisabilité et la fiabilité de notre approche, nous l'avons expérimentée sur un corpus de 97 histoires produites à l'oral par des enfants. Nous présentons le résultat de cette évaluation.

Abstract. We present an anaphora resolution integrated in a discourse analysis. This study deals with pronoun anaphora and zero anaphora. Our analysis is based on three approaches. A constraint based rule analysis, a functional approach and a dynamic analysis. In order to evaluate the feasibility of our approach and its reliability we have experimented with a corpus of 97 speech stories produced by children. We present here the results of our evaluation experiment.

Mots-clés : analyse de discours, résolution anaphorique, anaphore pronominales, anaphores zéro, grammaires de propriétés, grammaire fonctionnelle, analyse dynamique, discours oral.

Keywords: discourse analysis, anaphora resolution, pronoun anaphora, zero anaphora, property grammars, functional grammar, dynamic analysis, speech discourse.

1 Introduction

Nous traitons d'une résolution anaphorique intégrée au sein des autres traitements linguistiques. Plus généralement nous essayons d'intégrer au mieux les différentes analyses linguistiques au sein de l'analyse de discours. Cette résolution anaphorique s'inscrit dans un projet de réalisation d'un analyseur de discours pour classification automatique de textes (Nouali *et al.*, 2005). Pour réaliser cet analyseur, nous sommes confronté à la nécessité de résoudre certaines anaphores. Il s'agit des anaphores pronominales sujet et objet mais aussi des anaphores zéro. Notre présentation du projet met en exergue l'application des mêmes principes pour chaque niveau de l'analyse linguistique. Ces principes sont l'application de contraintes, une approche dynamique et la hiérarchisation de fonctions sémantiques, syntaxiques et pragmatiques. Dans un premier temps nous présentons l'application de ces principes pour le découpage en chunks et leurs rattachements, ainsi que pour l'assignation des fonctions, ensuite nous aborderons l'analyse anaphorique et les résultats obtenus.

Les sections vont s'articuler de la façon suivante. La section 2 dégage les propositions sur lesquelles nous avons axé notre analyse du discours. La section 3 décrit le fonctionnement de l'analyseur en présentant les trois types de mise à jour actuellement implémentée. La section 4 présente l'analyse utilisée pour la résolution des anaphores pronominales et des anaphores zéro. La section 5 présente les résultats.

2 Proposition

L'analyse que nous proposons est à la croisée de trois types d'approche. C'est une approche par contraintes et nous utilisons le paradigme des Grammaires de Propriétés tel qu'il est décrit dans (Blache, 2001). Nous utilisons aussi l'apport des constructions au sein des Grammaires de Propriétés (Blache, 2005). C'est aussi une approche fonctionnelle pour la place importante que nous accordons aux fonctions. Nous nous référons pour cela à la Grammaire Fonctionnelle de S. C. Dik (Dik, 1997). C'est une analyse de discours dynamique pour, à terme, être à même d'intégrer facilement des inférences sémantiques comme la SDRT le propose (Asher, 1993).

Nous appliquons les mêmes principes de contraintes, les principes fonctionnels et les principes de l'analyse dynamique pour tous les niveaux des traitements linguistiques. De ce fait, à chaque niveau, on applique des contraintes comme les contraintes de linéarité par exemple. On observe aussi un ordre prioritaire de fonctions. Ces fonctions sont les fonctions sémantiques, syntaxiques et pragmatiques définies en Grammaire Fonctionnelle. L'analyse est incrémentale, le contexte sélectionne les contraintes que l'on applique et chaque entité nouvelle tente de s'intégrer à la structure précédemment créée.

3 L'analyseur

L'analyseur de discours a été implémenté en O'CAML par Gilles Régnier selon les spécifications que nous avons établies. Il construit une représentation du discours au moyen de trois types de mise à jour. La première mise à jour construit les segments de texte qui correspondent aux chunks d'un chunk parser. La deuxième concerne les rattachements des chunks entre eux. Cette mise à jour leur attribue de fait un rôle. La troisième est une mise à jour fonctionnelle. Elle met à jour les fonctions sémantiques, une fonction syntaxique et des fonctions pragmatiques. Après ces mises à jour une phase de filtrage appelée inhibition est activée pour limiter les effets de l'explosion combinatoire des interprétations proposées.

Le lexique utilisé est le lexique du Laboratoire Parole et Langage d'Aix-en-Provence que l'on a enrichi d'informations sémantiques comme la structure des prédicats, des restrictions de sélection, et des indications sur la classe sémantique à laquelle appartient l'item lexical.

3.1 La mise à jour des chunks

La frontière des constituants est déterminée par l'évaluation de chaque frontière de mot. Chaque mot entrant est évalué et attribue un score sur sa frontière gauche et/ou sa frontière droite. Il peut aussi apparaître dans un contexte spécifique et être intégré à une construction qui modifie ces scores. La règle 1 est un exemple extrêmement simplifié de hiérarchie pour la frontière

gauche d'un syntagme nominal donné à titre d'illustration. Celui-ci indique que les éléments les plus susceptibles d'ouvrir un syntagme nominal sur leur gauche, sont par ordre décroissant, un déterminant puis un adjectif cardinal puis un adjectif ordinal puis un nom.

(1) Det > Card > Ord > N

L'analyse est dynamique. La détermination des frontières des chunks constitue des contextes. Les contraintes sont appliquées en fonction des contextes ainsi créés. Par exemple les contraintes de linéarité sont différentes en fonction du type de chunk sur lesquelles on les applique. Si l'on est au sein d'un syntagme nominal, les contraintes de linéarité spécifiques au syntagme nominal sont appliquées.

3.2 La mise à jour des rattachements

La mise à jour des rattachements relie les différents chunks entre eux. Lorsqu'un chunk est constitué, l'analyseur essaie de le rattacher à la structure. L'attachement est établi selon un ordre de priorité basé sur les fonctions linguistiques qui accorde une préférence sur certaines structures. Pour la linéarité des fonctions des arguments du verbe par exemple, l'ordre de préférence indique que le sujet précède (sauf cas particuliers) les compléments du verbe.

L'approche est dynamique. Chaque mise à jour réévalue l'ensemble des rattachements de la représentation du ou des prédicats en cours. Cependant seules certaines positions de la structure restent ouvertes au rattachement. Par exemple un pronom sujet qui précède un verbe verrouille le point de rattachement de la position sujet de ce verbe. La réévaluation des rattachements s'opère donc sur une structure qui est partiellement constituée.

On observe des contraintes sur la structure comme par exemple la contrainte de projectivité (les arcs des rattachements ne peuvent se croiser). On applique aussi des contraintes syntaxiques sur la ou les prépositions qui peuvent introduire la sous-catégorisation.

3.3 La mise à jour des fonctions

La mise à jour des fonctions assigne à certains termes les fonctions de la Grammaire Fonctionnelle. La fonction syntaxique *sujet* et les fonctions pragmatiques *Thème*, *Topique* et *Focus* sont centrales pour notre résolution anaphorique comme nous le verrons en section 4.

Les fonctions sémantiques assignent par exemple les rôles d'*agent*, de *patient* aux arguments des prédicats.

La mise à jour des fonctions syntaxiques ne concerne pour le français, conformément à la Grammaire Fonctionnelle de S. C. Dik, que la fonction *sujet*. Cette assignation est utilisée lors des rattachements. La fonction *Sujet* est aussi mise à contribution pour la résolution des anaphores zéro et des anaphores pronominales.

La mise à jour des fonctions pragmatiques assigne les fonctions de *Topique*, de *Focus* et de *Thème* de la Grammaire Fonctionnelle. La fonction pragmatique *Thème* est assignée au syntagme nominal des extractions à gauche. La fonction pragmatique *Focus* est assignée au syntagme nominal ou prépositionnel d'une construction clivée.

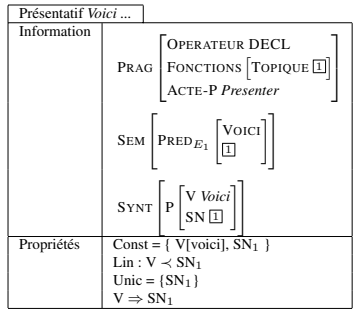


FIG. 1 – Construction du présentatif *Voici ...*

La fonction pragmatique *Topique* est assignée de façon très diverse. Elle peut être déterminée au moyen de constructions particulières. La figure 1 représente la construction *Voici* qui s'applique sur des phrases comme l'exemple 2.

(2) Voici un ver.

Lorsque les propriétés définies dans l'ensemble des propriétés sont satisfaites, les informations associées de la construction sont propagées dans la structure. On peut observer que les informations pragmatiques de cette construction assignent la fonction de *Topique* à un argument du verbe *Voici*. Celui-ci est un syntagme nominal précédé par *Voici*. La résolution anaphorique que nous proposons utilise ce type d'information pragmatique.

La fonction *Topique* peut aussi être assignée au syntagme nominal tête d'une phrase nominale. Dans la majeure partie des cas, en l'absence de construction particulière, elle est assignée au syntagme auquel est assignée la fonction syntaxique *Sujet*.

3.4 La phase d'inhibition et la sortie de l'analyseur

Au fur et à mesure de la construction des éléments et de l'application des contraintes une évaluation de la structure est faite. Deux types d'évaluation sont utilisés. Le premier concerne la linéarité et le deuxième concerne le nombre de rattachements. Ces évaluations sont indépendantes. Une interprétation est éliminée lorsque la différence entre son score et celui de l'interprétation la mieux évaluée dépasse un seuil fixé. Cette évaluation permet d'éliminer en cours d'analyse les interprétations que l'on estime les moins cohérentes.

Pour observer les résultats de l'analyseur un ensemble d'informations est extrait des objets construits et exprimé dans un graphe orienté. La figure 2 est une visualisation d'un graphe issu d'un des segments de texte de notre corpus. Les numéros qui précèdent chaque extrait de texte indiquent l'ordre d'apparition dans le segment de texte. Les chunks verbaux sont représentés par des rectangles. Les conjonctions de coordination sont représentées par des triangles. Les ellipses représentent les syntagmes nominaux et les syntagmes prépositionnels. Le cercle représente certains adverbes (comme ceux de lieu et temps) identifiés comme satellite (selon la terminologie de la Grammaire Fonctionnelle). L'élément tête de chaque chunk est représenté entre crochets.

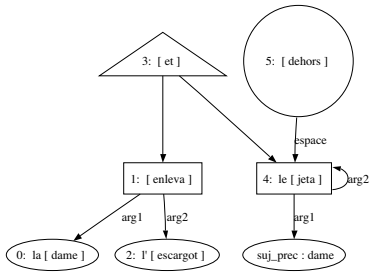


FIG. 2 – Sortie d'une interprétation de la phrase *La dame enleva l'escargot et le jeta dehors*

4 La résolution anaphorique

La résolution anaphorique concerne deux types d'anaphore : les anaphores pronominales et les anaphores zéro. Des études (dont (Ide & Cristea, 2000)) ont montré des relations entre la structure du discours (notamment RST, théorie des veines) et la résolution anaphorique. La structure de discours contraint l'accessibilité des référents. Or l'analyseur ne fournit pas beaucoup d'informations sur la structure rhétorique du discours. (Tetreault & Allen, 2004) ont mené une expérience sur un corpus de dialogues dont la structure de discours a été annotée. Ils en concluent que la prise en compte de la structure du discours n'apporte pas statistiquement d'améliorations significatives. La structure de discours n'étant que partiellement réalisée par notre analyseur, cela ne devrait pas générer un impact important sur nos résultats.

4.1 Les relations anaphoriques pronominales

Les anaphores pronominales traitées sont restreintes aux pronoms clitiques de la troisième personne ; *il, elle, ils, elles* en tant que pronom sujet et *le, la, lui, les, leur* en tant que pronom préverbal complément. La résolution ne concerne pas les anaphores abstraites (Amsili *et al.*, 2006).

Pour résoudre ces anaphores, on a utilisé les principes de la théorie du centrage (Grosz *et al.*, 1995). Cette théorie décrit le fonctionnement de la structure attentionnelle du discours qui contribue à la cohérence discursive. Il s'agit de la mise en correspondance de deux ordres de centres attentionnels : les centres anticipateurs et les centres rétrospectifs. Lorsque deux énoncés se succèdent le premier énoncé contient des centres attentionnels ordonnés du plus saillant au moins saillant. Les centres rétrospectifs vont se connecter sur les centres anticipateurs et ce plus favorablement vers les plus saillants. Les centres rétrospectifs les plus à même de se connecter sur les centres anticipateurs sont les pronoms anaphoriques. La théorie du centrage énonce des règles pour ordonner les centres attentionnels de sorte à pouvoir les apparier.

Dans notre cas, la hiérarchie des antécédents est basée sur les fonctions pragmatiques de la grammaire fonctionnelle (Dik, 1997) Part 1, Chap.13, pp309–338. L'ordre utilisé est donné dans l'exemple 3. Cette hiérarchie indique que l'élément auquel on a assigné la fonction *Thème* est prioritaire sur un élément auquel on a assigné la fonction *Focus*. Ce dernier est prioritaire sur un élément auquel on a assigné la fonction *Topique*. Les autres arguments sont ordonnés selon

leur ordre d'apparition dans la phrase.

(3) Fonction Thème > Fonction Focus > Fonction Topique > autres arguments

Les différentes fonctions sont identifiées par les constructions spécifiques comme indiqué dans la partie 3.3. L'identification d'une construction spécifique comme l'extraction à gauche permet d'assigner la fonction *Thème*. La fonction *Focus* est identifiée grâce aux constructions clivées. La fonction Topique est identifiée par des constructions diverses comme *Il était une fois ..., il y a ..., c'est un ...*.

Les centres rétrospectifs qui nous intéressent sont les pronoms. Nous traitons deux types de pronoms. Les pronoms sujet et les pronoms objet. Ces deux types de pronoms sont ordonnés dans le cas où les deux types sont en présence simultanément dans le même énoncé. Comme la règle 4 l'indique, nous favorisons les pronoms sujet que nous affectons en priorité à l'antécédent le plus saillant.

(4) pronoms sujet > pronoms complément

Dans la théorie du centrage, la cohérence est locale et les règles ne s'appliquent que sur deux phrases consécutives. (Alshawi, 1987) propose un ordre de saillance au moyen de pondérations. Ceci permet d'étendre les rapports entre pronoms et antécédents sur plusieurs phrases. Chaque apparition d'un référent de discours ajoute un nombre de points qui varie selon la position de ce référent dans la phrase. Ce capital de points associé au référent diminue au fur et à mesure des phrases s'il n'est pas cité. Une illustration de ce type de pondération est présentée dans (Lappin & Leass, 1994) et pour le français dans (Victorri, 2005).

Nous appliquons une méthode mixte qui d'une part utilise une hiérarchie attentionnelle entre les énoncés successifs et qui d'autre part pondère les référents du discours sur des empans plus larges que deux phrases successives. Nous mémorisons dans une pile les référents du discours selon un ordre de saillance dépendant de leur occurrence dans le discours. Cependant notre pondération ne dépend pas de l'étiquette syntaxique des éléments comme pour (Lappin & Leass, 1994). Elle dépend de l'ordre de saillance dans chaque énoncé. L'élément le plus saillant d'un énoncé reçoit toujours la même pondération quelle que soit sa position syntaxique. De plus notre pondération comprend des exceptions. Comme nous allons le voir plus loin, la fonction *Thème* manipule directement la hiérarchie des référents. De même une analogie structurelle entre deux énoncés successifs favorise la coréférence. Enfin nous accordons une pondération plus forte pour les référents qui apparaissent sous forme de pronom car c'est l'indice d'un ancrage plus fort dans le discours.

Nous avons utilisé la pondération suivante. Pour chaque énoncé, le premier élément de l'ordre des antécédents reçoit 100 points et 120 s'il s'agit d'un pronom. Le deuxième élément reçoit 50 points et 60 s'il s'agit d'un pronom. Le passage d'un énoncé au suivant a pour effet de multiplier tous les scores par 0,6. Si les exemples 5, 6 et 7 se suivent alors en 5, *Jean* a le score de 100. En 6, *Herbert* a le score de 100 et *Jean* qui est lié au pronom *le* a le score de 120 (soit 0,6x100+60). De ce fait, selon notre pondération, en 7, *Il* est attribué à *Jean*.

(5) Jean marche dans la rue.

(6) Herbert le salue.

(7) Il hoche la tête.

(8) Il l'interroge.

Cette affectation au moyen de pondération comprend des exceptions. C'est le cas lorsqu'un énoncé dont le verbe a deux arguments instanciés est suivi d'un énoncé de même structure dont les deux arguments sont pronominalisés. On attribue alors par analogie structurelle les arguments dans leur ordre respectif quelle que soit leur pondération. Par exemple si la phrase 8 succédait à 6, on attribue *Herbert* à *Il* et *Jean* à *l'* même si *Jean* a une plus forte pondération que *Herbert*.

Une autre exception concerne la fonction pragmatique de *Thème*. Cette fonction est prioritaire sur la saillance. Nous l'interprétons comme une modification de cette saillance. L'élément auquel est assignée la fonction *Thème* obtient le score de l'élément le plus élevé de la pile et se voit attribuer prioritairement le centre rétrospectif le plus élevé. Quel que soit le score des antécédents dans l'énoncé qui précède l'exemple 9, nous apparions *Le monsieur* à *il*.

(9) Le monsieur, il le prend.

D'autres règles spécifiques sont appliquées comme l'assignation de la fonction topique au complément du verbe, si le pronom "on" est sujet. On utilise des contraintes comme la contrainte d'accord en genre et en nombre. Cette contrainte est très importante. (Beaver, 2004) qui utilise la théorie de l'optimalité pour une résolution d'anaphore classe l'accord comme la contrainte de rang le plus fort.

Lorsque l'analyseur propose plusieurs analyses pour une même séquence, il faut faire un choix pour transmettre à la séquence de texte suivante les antécédents en cours. Afin de résoudre ce problème et permettre la transition, nous avons choisi de sélectionner la solution majoritaire proposée dans l'ensemble des interprétations. Nous avons utilisé le même procédé pour présenter les résultats. Au sein d'une séquence textuelle, si plusieurs interprétations sont proposées seul le résultat majoritaire est retenu.

4.2 Les relations anaphoriques zéro

Les anaphores zéro ne concernent que les sujets omis substituables par des pronoms clitiques sujet à la troisième personne. La résolution est simple et uniquement basée sur la fonction syntaxique *Sujet*. Lorsque le sujet est manquant dans un énoncé, on lui attribue le sujet de l'énoncé précédent en vérifiant les contraintes d'accord. La figure 2 illustre la résolution d'une anaphore zéro. L'analyseur attribue pour sujet du verbe *jeta* l'élément tête du sujet de la phrase qui précède *dame*.

5 Résultats

Le corpus utilisé est une partie d'un corpus constitué par Monique Vion et Annie Colas (Vion & Colas, 2000). Pour produire ce corpus, des bandes dessinées ont été présentées à des enfants. On a demandé aux enfants de raconter l'histoire que présentent ces images. Ce corpus est actuellement intégré au corpus CHILDES. Nous avons retenu pour l'évaluation une série de ces histoires appelée "Ver et escargot". D'autres séries nous ont servi de corpus d'observation et d'entraînement. Nous avons utilisé les transcriptions des enfants de 7, 9 et 11 ans. Cela représente un ensemble de 97 textes.

Une mise en forme des textes a été nécessaire. Les balises spécifiques de CHILDES ont été enlevées et certaines conventions de transcription ont été corrigées. Par exemple lorsque le pronom *il* était prononcé *i* le transcripteur l'a écrit *i*. Certaines disfluences ont été corrigées. Il s'agit de répétitions, nous n'avons conservé que la partie droite de ces répétitions. Avant d'effectuer l'analyse nous nous sommes assuré que tous les mots de notre corpus renvoient à une ou plusieurs entrées de notre lexique. Les ajouts étaient composés d'interjections et de noms propres.

Au sein de ces histoires nous avons identifié 671 anaphores pronominales et 31 anaphores zéro. Les textes ont été soumis à une évaluation de trois sujets. Pour chaque pronom et anaphore zéro, ils ont indiqué le nom de l'antécédent. Ils avaient le choix entre trois solutions. S'ils identifiaient l'antécédent, ils indiquaient le nom de l'antécédent. Dans le cas de pronoms pluriels à antécédents multiples, ils pouvaient les lier au moyen d'un signe plus. Si aucun antécédent n'était identifié ou si une ambiguïté des antécédents était manifeste, ils notaient un point d'interrogation. 10% des antécédents n'ont pas obtenu d'accord total entre les sujets quelle que soit la réponse pour les anaphores pronominales. Et 3% n'ont pas obtenu d'accord unanime pour les anaphores zéro.

Anaphores pronominales	Précision
Anaphores pronominales consensuelles	86 %
Anaphores pronominales non consensuelles	95%

TAB. 1 – Évaluation des anaphores pronominales

Sur les 671 anaphores pronominales, 18 n'ont pas été prises en compte puisqu'elles ne présentaient pas d'intérêt pour la suite de nos travaux. Ce sont par exemple des anaphores incluses dans une subordonnée relative. 35 anaphores supplémentaires n'ont pas été prises en compte puisqu'elles ont été jugées ambiguës ou sans antécédent par les sujets. Les résultats du tableau 1 indiquent la précision obtenue sur 616 anaphores. Les anaphores consensuelles sont celles dont les sujets ont désigné un antécédent à l'unanimité. Les non consensuelles sont celles où seule la majorité des sujets a désigné l'antécédent attendu. L'essentiel des erreurs a lieu dans des configurations où les relations entre les pronoms et leurs antécédents ne sont pas ambiguës.

Les résultats du tableau 2 portent sur l'ensemble des anaphores pronominales (consensuelles et non consensuelles confondues) et les anaphores zéro. Les résultats semblent satisfaisants lorsqu'on les compare à d'autres travaux sur la résolution anaphorique comme (Trouilleux, 2002) et (Tetreault, 2001). Peut-être, la nature du corpus facilite-t-elle notre tâche. Il s'agit pour nous d'un corpus oral alors que les exemples cités sont appliqués sur des corpus écrits. L'observation de ce corpus montre les textes oraux peuvent être difficiles pour la résolution anaphorique puisque 10% des textes présentent des difficultés de résolution pour les sujets interrogés. Il n'y a par exemple pas d'antécédent pour une partie des pronoms. Cependant lorsque la résolution est possible, il semble que l'information est présentée de façon plus conforme aux prédictions de la théorie du centrage que les textes écrits. A ce titre nous avons favorisé les anaphores interphrastiques et la continuité thématique comme la théorie du centrage le préconise. Cela nous semble plus approprié aux corpus oraux. On peut remarquer aussi que la couverture des analyses citées sur les textes écrits est aussi plus large puisqu'elles intègrent une résolution des adjectifs possessifs.

Type d'anaphore	Précision
Anaphores pronominales	87 %
Anaphores zéro	93%

TAB. 2 – Évaluation générale

6 Conclusion

L'intégration de différents niveaux linguistiques au sein de l'analyse de discours permet de profiter des informations des différents traitements pour résoudre des problèmes discursifs. L'analyse de discours met à disposition des informations fonctionnelles issues de constructions syntaxiques spécifiques. Nous avons pu observer que ces informations, comme la fonction sujet, peuvent remonter dans l'analyse et améliorer la résolution de problèmes discursifs comme la résolution anaphorique.

Les approches utilisées, les contraintes, les fonctions et l'approche dynamique se combinent au sein de notre analyse. Elles sont applicables à tous les niveaux d'analyse. Elles ne se confinent pas à un seul de ces niveaux.

L'observation des erreurs nous amène à quelques pistes d'amélioration de l'existant. Par exemple, la résolution des pronoms à antécédents multiples et la gestion de la cataphore n'étaient pas implémentées. Elles ont produit des erreurs. On peut envisager des améliorations en tenant compte de la diversité des fonctions pragmatiques. On pourrait affiner la pondération de la saillance des référents en exploitant mieux les différentes fonctions. On peut continuer à améliorer l'analyse de discours pour contraindre les énoncés afin de diminuer le nombre d'interprétation et simplifier la détermination des résultats.

D'autres erreurs demanderaient un investissement considérable notamment pour les résolutions qui font appel à notre connaissance du monde. Dans la suite des énoncés 10 et 11 l'antécédent de *Il* en 11 nous semble être plus favorablement relié au patient plutôt qu'à l'agent du verbe jeter. Et cela quel que soit le poids de *Il* en 10. Cette intuition fait appel à notre connaissance du monde.

(10) Il le jette.

(11) Il tombe.

D'autres problèmes liés à la résolution anaphorique pourraient se poser si nous changions de format ou de genre de texte. Par exemple les textes analysés ont en moyenne une douzaine de séquences de texte. Si les textes étaient plus longs, la pondération d'un référent du discours pourrait augmenter considérablement et ne plus permettre de transition. Nous devrions prévoir une valeur plafond que la pondération ne pourrait pas dépasser.

Références

ALSHAWI H. (1987). *Memory and context for language interpretation*. New York : Cambridge University Press.

AMSILI P., DENIS P. & ROUSSARIE L. (2006). Anaphores abstraites en français : représentation formelle. *Traitement Automatique des Langues*, **46**(1), 15–39.

ASHER N. (1993). *Reference to abstract objects in discourse*. Dordrecht : Kluwer Academic Publishers.

BEAVER D. (2004). The optimization of discourse anaphora. *Linguistics and philosophy*, **27**(1), 3–56.

BLACHE P. (2001). *Les Grammaires de Propriétés : des contraintes pour le traitement automatique des langues naturelles*. Paris : Hermès Sciences.

BLACHE P. (2005). Property grammars : A fully constraint-based theory. In H. CHRISTIANSEN & AL., Eds., *Constraint Solving and Language Processing*. Springer.

DIK S. C. (1997). *The theory of functional grammar. 2 Volumes. Part 1 : The Structure of the Clause. Part 2 : Complex and Derived Constructions*. Berlin : Mouton de Gruyter.

GROSZ B., JOSHI A. & WEINSTEIN S. (1995). Centering : a framework for modelling the local coherence of discourse. *Computational Linguistics*, **21**(2), 203–226.

IDE N. & CRISTEA D. (2000). A hierarchical account of referential accessibility. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL 2000*, p. 416–424, Hong-Kong.

LAPPIN S. & LEASS H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4), 535–561.

NOUALI O., RÉGNIER A. & BLACHE P. (2005). Classification de courriers électroniques : une approche par apprentissage basée sur des modèles linguistiques. *Revue d'intelligence artificielle*, **19**(6), 885–912.

TETREAULT J. R. (2001). A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, **27**(4), 27–4.

TETREAULT J. R. & ALLEN J. F. (2004). Dialogue structure and pronoun resolution. In COLIBRI, Ed., *Proceedings of DAARC 2004*, p. 7–12, Lisbonne.

TROUILLEUX F. (2002). A rule based pronoun resolution system for french. In COLIBRI, Ed., *Proceedings of DAARC 2002*, Lisbonne.

VICTORRI B. (2005). Le calcul de la référence. In HERMÈS, Ed., *Sémantique et traitement automatique du langage naturel*, p. 133–172, Paris.

VION M. & COLAS A. (2000). Mode de recueil et outil d'analyse d'un corpus de parole spontanée étudié d'un point de vue psycholinguistique. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence*, **19**, 155–167.

Atelier TALN-2007

5 au 8 juin 2007, Toulouse, France

Actes de l'atelier

FORMALISMES SYNTAXIQUES

DE HAUT NIVEAU

Éditeur scientifique

Lionel CLÉMENT

Sous l'égide de l'ARC-MOSAIQUE

(Action de Recherche Coopérative)

Comité d'organisation

Lionel CLÉMENT*
Marie-Laure GUÉNOT
Renaud MARLET
Christian RETORÉ
Benoît SAGOT
Tristan VANRULLEN

Comité de programme

Lionel CLÉMENT*
Denis BECHET
Philippe BLACHE
Benoît CRABBÉ
Bertrand GAIFFE
Claire GARDENT
Sylvain KAHANE
Éric DE LA CLERGERIE
Renaud MARLET
Guy PERRIER
Christian RETORÉ
Azim ROUSSANALY
Benoît SAGOT
Isabelle TELLIER

* Président

Session
Communications orales

Problématique de la conception d'un langage de haut niveau

Benoit CRABBÉ
LATTICE-CNRS et Université Paris 7
benoit.crabbe@linguist.jussieu.fr

Résumé. Cet article identifie les buts d'un langage de représentation grammaticale dit de haut niveau. Ceux-ci étant fixés il identifie les problèmes à adresser lors de la conception d'un langage de ce type. Il montre enfin comment ces buts et ces problèmes sont traités dans différents formalismes de représentation grammaticale. À l'issue de cette comparaison, nous distinguons un ensemble de problèmes noyaux traités dans ces langages d'un ensemble de problèmes secondaires spécifiques à certains formalismes.

Abstract. This paper identifies the goals of a computer language for grammatical representation. Those being given, the paper identifies the main problems to address when designing a language of this kind. It finally shows how these goals and these problems are managed in several grammatical description formalisms. The outcome of this comparison leads to a distinction between kernel problems to adress in the design of such a language and secondary problems that are formalism specific.

Mots-clés : formalismes d'analyse syntaxique, langage de représentation grammaticale, métagrammaire.

Keywords: grammatical formalisms, metagrammar.

1 Introduction

Le problème de l'analyse syntaxique automatique des langues naturelles ne se réduit pas à créer un algorithme d'analyse syntaxique et de l'utiliser tel quel. En effet, lorsqu'on s'attaque aux problèmes de conception de grammaires de taille importante, il devient rapidement crucial pour le développeur de grammaire de disposer d'un langage informatique de représentation grammatical facilitant l'expression manuelle de cette grammaire.

La raison est que les algorithmes d'analyse syntaxique classiques manipulent des systèmes formels dotés de propriétés formelles et calculatoires satisfaisantes à des fins de traitement automatique : ils sont décidables et la complexité en temps est généralement polynomiale. Il s'agit par exemple de grammaires de réécriture généralement augmentées de structures de traits[1].

Bien qu'adaptée à des fins de traitement automatique, la représentation d'une grammaire utilisée par un analyseur syntaxique n'est toutefois pas une représentation qui convient directement à

[1]Certaines implémentations autorisent l'usage de structures de traits récursives sans implémenter un « occur check », trop coûteux en temps, ce qui les rend théoriquement indécidables. Voir par exemple (Shieber *et al.*, 1995) pour une présentation des techniques d'implémentation actuelles.

un développeur de grammaire humain et en particulier aux développeurs linguistes. Celui-ci souhaite exprimer une grammaire en capturant des généralisations théoriques qu'il n'est pas possible d'exprimer avec une représentation d'aussi bas niveau (Gazdar *et al.*, 1985).

Pour cette raison, la plupart des systèmes d'analyse syntaxique d'envergure doivent proposer un langage de représentation grammaticale de haut niveau qui permet au grammairien d'exprimer une grammaire en manipulant des généralisations qui lui sont familières.

Ainsi la plupart des grammaires de grande taille sont exprimées dans un langage de haut niveau manipulé par le grammairien. Celui-ci est compilé dans un langage de bas niveau qui est compatible avec les propriétés informatiques des analyseurs syntaxiques.

LANGAGE DE HAUT NIVEAU
↓
Traduction
↓
LANGAGE DE BAS NIVEAU

Ce papier cherche à identifier les caractéristiques d'un langage de haut niveau ainsi que les problèmes qu'il s'agit de traiter lors de la conception d'un langage de ce type. Dans ce qui suit nous appellons *grammaire source* la grammaire décrite dans un langage de haut niveau et *grammaire cible* la grammaire exprimée dans le langage de bas niveau. L'article est structuré en trois parties : tout d'abord nous identifions des buts pour ce type de langages (Section 2). Ceux-ci étant fixés, nous identifions en section 3 les problèmes principaux qu'il s'agit de traiter lorsqu'on propose un langage de ce type, et ce quel que soit le formalisme envisagé. Finalement nous distinguons ces problèmes généraux de problèmes plutôt spécifiques aux formalismes contemporains en section 4.

2 Buts de la représentation grammaticale

Factorisation de l'information. Les langages de haut niveau sont utilisés dans une perspective appliquée afin de décrire des grammaires dites à large couverture. De ce point de vue ces langages de haut niveau ont pour fonction essentielle de permettre aux développeurs d'exprimer une grammaire de manière non redondante. Cela s'illustre en pratique en autorisant l'expression de l'information dans des modules, ou classes réutilisables et qu'il est possible de combiner entre-eux.

Capturer les alternatives. Ceci dit, les langages de représentation grammaticale ont d'abord été conçus pour tester et vérifier en pratique les théories et les descriptions grammaticales dérivées de la théorie. En ce sens ils se doivent de proposer aux linguistes les moyens d'exprimer la théorie de la manière la plus fidèle possible. Cette idée est initialement rendue explicite par (Gazdar *et al.*, 1985) : les auteurs montrent comment approximer une contrepartie satisfaisante d'une grammaire générative et transformationelle à l'aide d'une simple grammaire de réécriture libre de contexte par précompilation d'un langage de haut niveau, appelé *métagrammaire*[2].

[2]Ceci dit, il ne faut pas oublier que GPSG est un formalisme de descriptions surfacique de la langue dans lequel il n'est pas question de distinguer une structure profonde d'une structure de surface. En particulier la mé-

En particulier, les langages de haut niveau se distinguent essentiellement des langages de bas niveau manipulés par les analyseurs car ils permettent l'expression d'informations qu'il n'est pas possible de traiter à l'aide d'algorithmes d'analyse syntaxique classique, en particulier l'expression d'alternatives — comme l'alternative actif / passif — initialement formulées par des règles lexicales (Bresnan & Kaplan, 1982) ou des métarègles (Gazdar *et al.*, 1985) dans les premières grammaires d'unification.

Contrairement à des règles de factorisation ou à des règles de réécriture, les alternatives ont un statut formel crucial dans la formalisation en syntaxe car celles-ci contribuent à mettre en relation un ensemble de structures exprimant la paraphrase. Ainsi l'expression d'une alternative actif/passif contribue à définir une relation entre *deux* réalisations d'un même mot. Celle-ci a donc un statut formel radicalement différent de règles de réécriture qui portent sur la définition de la grammaticalité des phrases du langage.

Extensibilité. En plus de la factorisation et du partage d'alternatives, il est utile d'introduire dans ce type de langage un certain nombre de notions théoriques supplémentaires, celles-ci sont dépendantes de la théorie sous-jacente à l'implémentation. Par exemple dans le cas d'une grammaire de type GPSG/HPSG, il est souhaitable d'avoir des mécanismes supplémentaires qui permettent de faciliter l'expression de la propagation de traits de tête, de la valence ou la gestion des dépendances à longue distance.

Spécificité (Lexicalisme fort). Comme illustré clairement par le premier formalisme explicite de ce type PATR II (Shieber, 1984), les langages de description grammaticale traditionnels comportent deux aspects : (1) Représentation de la grammaire et de ses règles et (2) la représentation du lexique. Ces vingt dernières années, les évolutions en linguistique théorique et formelle ont toutefois déplacé l'essentiel des préoccupations en syntaxe sur les aspects lexicaux.

Dans ce qui suit, nous nous limitons à la problématique de la définition d'un langage de haut niveau pour formalismes fortement lexicalisés. Autrement dit, nous traitons de formalismes qui n'opèrent pas formellement de distinction nette entre le lexique et la grammaire, comme par exemple les grammaires d'arbres adjoints ou les grammaires catégorielles[3]. Il est toutefois possible à partir de langages de ce type, comme le montre par exemple (Clément & Kinyon, 2003) d'engendrer des grammaires cibles pour des formalismes faiblement lexicalisés, comme LFG.

Dans ce qui suit, nous illustrons notre propos par un exemple simplifié issu de la grammaire TAG que nous avons développé (Figure 1). Nous considérons que cet exemple illustre la représentation de bas niveau produite en résultat de compilation d'un langage de haut niveau

3 Problèmes généraux

Cette section propose d'identifier trois problèmes de conception qui sont communs à la plupart des formalismes de représentation grammaticale connus : **la factorisation d'information, la**

tagrammaire GPSG possède des propriétés formelles totalement différentes d'une grammaire transformationelle : les métarègles qui la composent sont précompilées "offline" ; celles-ci ont une **localité** réduite : elles portent sur des règles de grammaires de profondeur 1 alors que les transformations portent potentiellement sur la totalité de l'arbre syntagmatique.

[3] Généralement les implémentations de ces formalismes rétablissent une distinction de ce type.

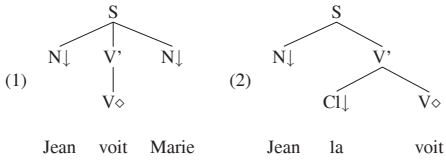


FIG. 1 – Extrait élémentaire de grammaire TAG

composition d'information et **l'expression d'alternatives**. Nous illustrons chacun des aspects à l'aide d'exemples inspirés de la grammaire d'arbres adjoints et du langage XMG qui manipule des descriptions de données arborescentes. Ceci dit, il est remarquable de constater que chacune de ces problématiques est adressée dans d'autres langages de représentation grammaticales utilisés pour traiter d'autres formalismes faiblement ou fortement lexicalisés qui manipulent des structures de données différentes. En particulier le langage de XLE (Dalrymple *et al.*, 2004) (structures de traits, LFG) et un langage HPSG qui manipule des structures de traits typées (Cohen-Sygal & Wintner, 2006).

Factorisation de l'information. Les informations exprimées dans une représentation formelle de bas niveau sont fortement redondantes. Un des premiers buts d'un langage de haut niveau est de permettre la factorisation d'information, c'est à dire à *nommer* des structures syntaxiques partielles qu'il est possible de réutiliser par ailleurs pour décrire des structures complètement spécifiées. L'exemple suivant illustre de manière abstraite comment sont nommés des fragments d'arbres dans XMG : les noms sont indiqués à gauche de la flèche pointant sur les fragments d'arbres auxquels ils sont associés : Ces descriptions partielles sont généralement ap-

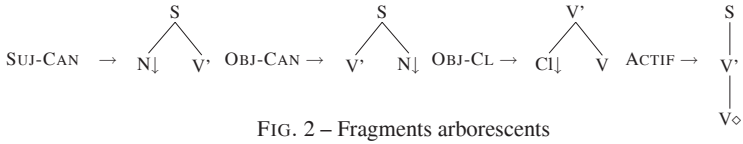


FIG. 2 – Fragments arborescents

pelées macros (Dalrymple *et al.*, 2004; Shieber, 1984), classes (Candito, 1996; Crabbé, 2005) ou modules (Cohen-Sygal & Wintner, 2006). Nous verrons par la suite que dans les systèmes XMG et HPSG de (Cohen-Sygal & Wintner, 2006) les classes ou modules définissent un espace de nom, ce qui n'est pas le cas des systèmes à macros, comme PATR II (Shieber, 1984).

Composition de modules. Les descriptions partielles ayant été définies, il faut fournir une opération de composition de ces descriptions. Cette opération doit être associative et commutative de manière à éviter des problèmes d'ordonnancement. Dans les systèmes à structure de traits non typées, il s'agit de l'unification de descriptions partielles (Shieber, 1984; Dalrymple *et al.*, 2004). Dans les systèmes à base de descriptions d'arbres (XMG) il s'agit d'une conjonction de formules décrivant la structure arborescente.

Expression des alternatives. L'intégralité des langages de représentation de haut niveau permettent d'exprimer des alternatives lexicales. Il s'agit de représenter qu'une même unité lexicale

peut se réaliser en syntaxe de manières différentes, pour un verbe on veut par exemple exprimer des alternatives comme l'actif et le passif. Trois solutions à ce problème sont envisagées : approche par règles lexicales, approche par covariance, approche par sous-spécification.

L'*approche par règles lexicales* suppose de distinguer des unités lexicales de base ou canoniques, d'unités lexicales dérivées. Les secondes sont engendrées à partir des premières à l'aide d'un mécanisme de règles lexicales (parfois appelées improprement métarègles). Une règle lexicale est une règle de la forme :

PARTIE GAUCHE $\longrightarrow$ PARTIE DROITE

Où la partie gauche spécifie un motif à reconnaître dans une entrée lexicale, la partie droite indique la transformation à opérer sur le motif reconnu pour produire une nouvelle entrée lexicale. La transformation permet d'ajouter, de supprimer ou de revoir de l'information.

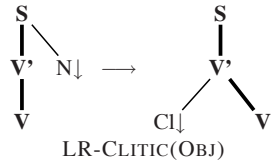


FIG. 3 – Règle lexicale de cliticisation

En considérant l'arbre de gauche en Figure 1 comme un arbre canonique, la règle lexicale donnée en Figure 3 engendre la variante de droite en Figure 1. Plus généralement le processus de dérivation est le suivant. Soit une liste de règles lexicales $L = l_1, l_2 \ldots l_n$ Le processus de dérivation d'une liste $A = a_1, \ldots a_n$ d'alternatives lexicales à partir d'une entrée de base a_1 est le suivant. Soit D une liste auxiliaire qui sert au calcul de la dérivation. D contient initialement d_1, l'entrée lexicale de base. Pour chaque règle lexicale $l_i \in L$, si la partie gauche de l_i est compatible avec d_1[4] alors produire une nouvelle entrée lexicale d_n qui est le résultat de l'application de la transformation spécifiée par l_i sur d_1 et ajouter d_n à D. Supprimer d_1 de D et l'ajouter à A. Le processus de dérivation termine lorsque la liste D est vide.

Le processus de dérivation n'est pas garanti de terminer, rien n'empêche de définir des règles lexicales ou des ensembles de règles lexicales qui créent des boucles. Le cas de figure le plus simple est celui de la règle lexicale $X \rightarrow X$ qui est la règle qui recopie son entrée vers sa sortie.

Outre la non garantie de terminaison de la procédure, le mécanisme à règles lexicales a été abondamment critiqué dans les années 90, car il induit une mécanique procédurale. La spécification d'une grammaire à large couverture à l'aide de ce genre de mécanique pose en pratique de sérieux problèmes pour gérer l'ordre d'application des règles.

Nous présentons maintenant deux manières de reformuler le problème des alternatives de manière déclarative. L'*approche par covariantes* est une alternative à l'approche par règles lexicales. Celle-ci s'en démarque fondamentalement dans la mesure où l'on ne distingue pas d'unité canonique des unités dérivées. Décrire une unité lexicale L revient à décrire un ensemble d'alternatives $a_1 \vee \ldots \vee a_n$ disjointes.

[4]Dans un système à structures de traits « compatible » signifie que les deux structures sont unifiables. (Becker, 1993) définit explicitement une contrepartie pour TAG.

Ainsi dans le système XMG on décrira les alternatives schématisées en Figure 1 par la description grammaticale (simplifiée) suivante :

VERBE TRANSITIF → SUJ-CAN ∧ ACTIF ∧ OBJET

OBJET → OBJ-CAN ∨ OBJ-CL

Le choix est interprété de manière indéterministe ainsi l'évaluation de la description VERBE TRANSITIF sera expansée par l'interprète en deux formes conjonctives :

VERBE TRANSITIF → SUJ-CAN ∧ ACTIF ∧ OBJ-CAN

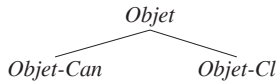
VERBE TRANSITIF → SUJ-CAN ∧ ACTIF ∧ OBJ-CL

Pour chacune de ces formes conjonctive, un interprète XMG construit les structures arborescentes en substituant les représentations arborescentes données en figure 2 dans les formules puis en composant les descriptions arborescentes.

L'approche par covariantes se distingue de l'approche à règles lexicales par son aspect déclaratif et le processus de dérivation des entrées est garanti de terminer. Ce procédé est utilisé dans deux langages contemporains ((Dalrymple *et al.*, 2004),(Crabbé, 2005)) ainsi que dans le langage de description du système HPSG de (Meurers & Minnen, 1997).[5]

L'*approche par sous-spécification* Il s'agit d'une approche utilisée principalement dans les grammaires HPSG. Celle-ci repose sur la sous-spécification de types. Tout objet grammatical est associé à un type. Les types sont partiellement ordonnés en semi-treillis, il s'agit de la hiérarchie d'héritage. Un type maximal est un type qui n'a pas de successeur. Lors de l'instanciation d'un objet grammatical, celui se voit assigner de manière indéterministe un type maximal.

Cet méthode a été réutilisée pour TAG par (Candito, 1996). Ainsi elle organise explicitement une hiérarchie d'héritage de classes comportant des fragments arborescents. Elle exprime ainsi les alternatives de réalisation fonctionelle par sous-spécification. Voyons à titre d'exemple un extrait de sa hiérarchie :



En assignant à un objet grammatical le type sous-spécifié *Objet*, la procédure d'instanciation de cet objet lui attribue de manière indéterministe un type maximal qui dans ce cas-ci est au choix *Objet-Can* ou *Objet-Clitique*.

Il est remarquable de constater que (Candito, 1999) et (Crabbé, 2005) (1) assignent les mêmes descriptions structurelles à des descriptions comme *Objet-Can* ou *Objet-Cl* (structures données en Figure 2) et (2) que les descriptions partielles définies par (Candito, 1999) et (Crabbé, 2005) correspondent aux parties affectées par les transformations structurelles de règles lexicales (voir Figure 3). Plus encore, la description par covariantes suivante :

OBJET → OBJET-CAN ∨ OBJET-CL

typiquement exprimée dans XMG (Crabbé, 2005) est non seulement une contrepartie de la description grammaticale par héritage donnée ci-dessus mais l'une et l'autre constituent également une contrepartie de la règle lexicale donnée en Figure 3.

[5] On peut par ailleurs conjecturer qu'il est formellement possible de formuler une traduction d'un système à covariantes dans un système à règles lexicales, l'inverse n'étant bien entendu pas vrai.

Cependant, à voir la multiplicité des solutions, la représentation des alternatives est très clairement un point délicat pour les langages de haut niveau. La différence principale entre les systèmes à covariantes et à sous-spécification d'une part et le système à règles lexicales d'autre part est le caractère déclaratif des deux premiers, et procédural du dernier.

Proposer un langage uniquement déclaratif ne va pas de soi. Faut-il encore trouver une méthodologie pour écrire la grammaire. Par exemple le langage du LKB (Copestake, 2002) et le système TAG de (Xia, 2001) proposent deux mécanismes pour capturer les alternatives : des règles lexicales et un système de sous-spécification de types. Dans ces systèmes, les règles lexicales sont utilisées pour modéliser des alternatives de diathèse, alors que la sous-spécification de type est utilisée pour modéliser des alternatives de réalisation de fonctions syntaxiques. La raison est que les alternatives de diathèse demandent d'exprimer une notion d'effacement. Il y a bien entendu le cas des passifs courts du type *Jean est vu* où l'agent n'est pas exprimé mais fait bien partie de la structure argumentale. Ce type de cas peut s'exprimer par sous-spécification dans un système déclaratif à partir du moment où on distingue deux niveaux de représentation : une représentation sémantique dans laquelle aucun argument n'est effacé et un niveau de représentation syntaxique dans laquelle le syntagme prépositionnel introduit par *par* est omis dans la réalisation syntaxique (Crabbé, 2005). Cela devient plus délicat si on veut considérer des cas comme l'alternance neutre *Jean casse une tasse* $\sim$ *Les tasses cassent*. Dans le second exemple, il est délicat d'identifier un second argument en sémantique. Cela signifie qu'on est en présence d'un cas où l'on cherche à traiter l'effacement d'un argument en sémantique. Qu'il s'agisse de formalismes à un seul niveau de représentation ou de grammaires qui cherchent à modéliser des cas analogues au neutre, l'un ou l'autre de ces deux aspects explique que les règles lexicales sont encore parfois utilisées dans les langages de haut niveau[6].

4 Problèmes spécifiques

Selon les formalismes et les grammaires, des problèmes supplémentaires doivent être pris en compte. Trois grandes catégories de problèmes sont identifiés : (1) Problèmes de nommage et (2) Problèmes d'interaction de descriptions et contraintes additionelles d'extensibilité.

Problèmes de nommage. Si le langage de XLE, et la description des grammaires LFG qui lui sont associées, permet de manipuler des structures de traits dont les noeuds sont anonymes[7], les langages conçus pour des grammaires accordant plus d'importance au lexique comme les langages HPSG de (Cohen-Sygal & Wintner, 2006) et TAG de (Crabbé & Duchier, 2004; Thomasset & Villemonte de La Clergerie, 2005) doivent traiter de problèmes de nommage.

Nous commençons par illustrer les problèmes de nommage des structures par un exemple extrait des grammaires TAG (Figure 4). La question à traiter consiste à déterminer comment combiner les fragments factorisés entre eux.

On envisage successivement trois manières de procéder. La première méthode part du principe que les fragments d'arbres combinés doivent mener à la construction d'une unité grammaticale

[6] C'est également pour modéliser ce type de phénomènes que (Candito, 1999) introduit des mécanismes procéduraux de manière à pouvoir exprimer un effacement.

[7] En interprétant les structures de traits comme des graphes acycliques orientés (DAGS), les noeuds dont il est question ici sont les noeuds de ces graphes.

qui est un arbre doté d'une racine unique et dont les labels de noeuds représentant les catégories syntaxiques doivent être compatibles (Figure 4, première ligne). Cette méthode est celle où les noms des noeuds proprement dits sont laissés anonymes. Cependant cette méthode souffre d'un problème de surgénération, le premier modèle obtenu n'est pas souhaité. Autrement dit, on manque de contrôle sur la manière de combiner les fragments.

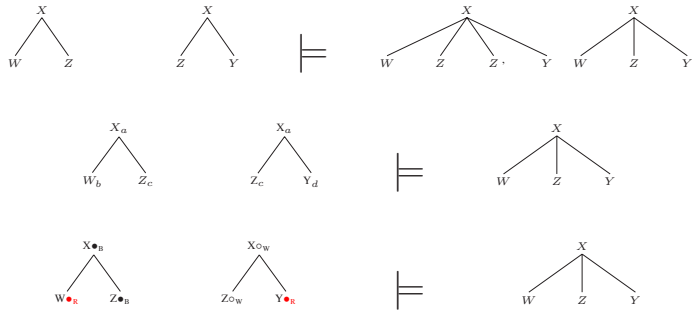


FIG. 4 – Problèmes de nommage

Une seconde méthode pour procéder consiste à contraindre d'avantage l'identification des noeuds en accordant un statut formel à leurs noms. C'est ce que fait (Candito, 1999) (Figure 4, seconde ligne). La contrainte supplémentaire qui est introduite lors de la recherche des modèles est que les noeuds de même noms doivent être identifiés[8]. Cette méthode donne un très grand contrôle sur la manière d'opérer les combinaisons mais présente l'inconvénient que l'ensemble des noms sont partagés par l'ensemble des classes décrivant une grammaire. Autrement dit il existe dans ce cas un seul espace de noms, commun à l'ensemble de la grammaire. On perd par là largement les bénéfices de la factorisation.

Une troisième méthode consiste à donner dans le langage des contraintes qui permettent de dénoter en intension des classes de noeuds (Figure 4, dernière ligne). Cette méthode est celle qui est introduite par (Crabbé & Duchier, 2004) en autorisant la décoration des noeuds à l'aide d'un jeu fini de couleurs (rouge, blanc, noir). Cette technique pose comme contrainte additionelle par rapport à la première méthode que chaque noeud blanc doit être saturé par un noeud noir. Cette méthode permet d'ajouter d'avantage de souplesse au langage de représentation grammatical sans toutefois imposer un espace de nom global à l'ensemble de la grammaire. On note toutefois que la solution colorée n'est pas pleinement satisfaisante dans la mesure où le langage XMG inclut également un mécanisme de gestion explicite d'espace de noms destinés à modéliser des cas représentant une notion d'héritage (Gardent & Parmentier, 2006). Dans ce sens, (Cohen-Sygal & Wintner, 2006) confrontés à un problème analogue dans leur langage de représentation HPSG utilisent pour résoudre le problème un système de types paramétrable qui permet à la fois de dénoter des noeuds extensionellement et à la fois intentionellement.

Contraintes additionnelles. Notons finalement que selon les formalismes, les différents langages de représentation autorisent l'usage de contraintes additionelles, ce que les auteurs de XMG appellent extensibilité. Ces contraintes additionelles ont un statut peu clair. Il peut s'agir

[8] Une variante de cette méthode est également proposée par (Vijay-Shanker & Schabes, 1992). Le langage qu'ils proposent autorise à contraindre explicitement l'identification de noeuds dotés de noms différents.

de contraintes visant à pallier les faiblesses du formalisme voire de certains choix d'implémentation. Nous illustrons ce point en Figure 5. Cette figure illustre la manière dont sont traités les

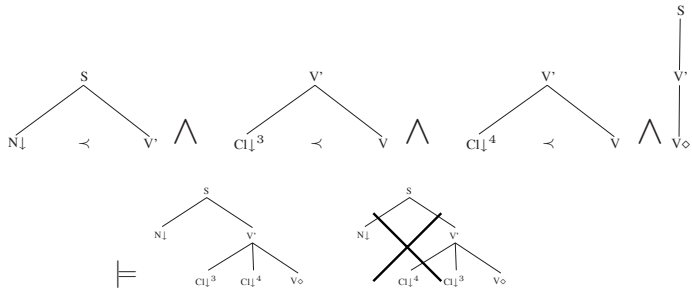


FIG. 5 – Interaction de descriptions

clitiques en XMG. Chaque argument clitique est décrit dans une classe indépendante : ainsi le critique de rang 3 *le* est décrit indépendamment du clitique de rang 4 *lui*. Chacune de ces descritpions indique (indirectement) que le clitique précède le verbe marqué en temps ou l'infinitif. Lorsqu'on engendre des structures impliquant deux clitiques, rien n'indique l'ordre de succession des clitiques relativement l'un à l'autre. Ce problème formel est relativement ancien, ainsi (Crabbé, 2005) implémente une solution inspirée de (Perlmutter, 1970) qui était confronté à un problème analogue en grammaire générative. Cette solution consiste à imposer une contrainte générale de bonne formation sur les arbres engendrés par une métagrammaire, stipulant que les clitiques fils d'un même noeud doivent être ordonnés selon leur propriété de rang.

5 Conclusion

Au terme de cette comparaison entre différents langages de représentation de haut niveau on constate que ceux-ci sont tous conçus pour faciliter l'expression de la grammaire sur les trois points suivants : (1) mise en facteur de l'information et nommage de descriptions grammaticales partielles (2) Réutilisation d'information et combinaison d'informations partielles et (3) expression d'alternatives.

Ce qui distingue les langages contemporains de langages plus anciens comme PATR II tient principalement dans la manière de traiter les alternatives. Les langages contemporains utilisent massivement des mécanismes purement déclaratifs en lieu et place de règles lexicales. Cependant dans bon nombre d'entre-eux, les règles lexicales restent encore parfois utilisées pour résoudre des problèmes formels spécifiques.

Les langages de représentation grammaticale contemporains se distinguent également de PATR II qui permet uniquement la manipulation de structures de traits par la manipulation de structures de données plus complexes : comme des arbres ou des structures de traits typées. La manipulation de structures plus complexes dans des formalismes portant une attention accrue à la description lexicale (TAG ou HPSG) apporte de nouveaux problèmes formels : comme typiquement la gestion d'espaces de noms ou la mise en évidence de problèmes d'interactions entre modules spécifiés indépendamment. À l'heure actuelle, pour traiter ce type de problèmes,

la tendance va vers l'emprunt de méthodes formelles issues des acquis en informatique dans le domaine de la conception de langages de programmation (Cohen-Sygal & Wintner, 2006).

Références

BECKER T. (1993). *HyTAG : A new Type of Tree Adjoining Grammars for Hybrid Syntactic Representation of Free Word Order Language*. PhD thesis, Universitat des Saarlandes.

BRESNAN J. & KAPLAN R. M. (1982). *The Mental Representation of Grammatical Relations*. Cambridge MA : The MIT Press.

CANDITO M.-H. (1996). A principle based hierarchical representation of LTAGs. In *COLING 96*, Copenhagen.

CANDITO M.-H. (1999). *Organisation Modulaire et Paramétrable de Grammaires Electroniques Lexicalisées*. PhD thesis, Université de Paris 7.

CLÉMENT L. & KINYON A. (2003). Generating LFGS with a metagrammar. In *Proc. LFG-03*, Saratoga Springs.

COHEN-SYGAL Y. & WINTNER S. (2006). Partially specified signatures : A vehicle for grammar modularity. In *Proceedings of the 44th meeting of the Association for Computational Linguistics*.

COPESTAKE A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI publications.

CRABBÉ B. (2005). Grammatical development with XMG. In P. BLACHE & E. STABLER, Eds., *Proceedings of Logical Aspects of Computational Linguistics*, Bordeaux.

CRABBÉ B. & DUCHIER D. (2004). Metagrammar redux. In H. CHRISTIAENSEN, Ed., *Constraint Solving and Language Processing*, Copenhagen.

DALRYMPLE M., KAPLAN R. & KING T. H. (2004). Linguistic generalizations over descriptions. In *LFG 2004*, Christchurch.

GARDENT C. & PARMENTIER Y. (2006). Coreference handling in xmg. In *COLING*, Sydney.

GAZDAR G., KLEIN E., PULLUM G. & SAG I. (1985). *Generalized Phrase Structure Grammar*. Harvard Univesity Press.

MEURERS W. D. & MINNEN G. (1997). A computational treatment of lexical rules in hpsg as covariation in lexical entries. *Computational Linguistics*, **23**(4), 543–568.

PERLMUTTER D. (1970). Surface structure constraints in syntax. *Linguistic Inquiry*, **1**, 187–255.

SHIEBER S. M. (1984). The design of a computer language for linguistic information. *COLING-84*, p. 362–366.

SHIEBER S. M., SCHABES Y. & PEREIRA F. (1995). Principles and implementation of deductive parsing. *Journal of Logic Programming*, **24**(1-2), 3–36.

THOMASSET F. & VILLEMONTE DE LA CLERGERIE E. (2005). Comment obtenir plus des méta-grammaires. In *Proceedings of TALN'05*, Dourdan, France : ATALA.

VIJAY-SHANKER K. & SCHABES Y. (1992). Structure sharing in lexicalized tree-adjoining grammars. In *COLING 92*.

XIA F. (2001). *Automatic Grammar Generation from two Different Perspectives*. PhD thesis, University of Pennsylvania.

Pour une représentation décentralisée de l'information syntaxique

Philippe BLACHE
Aix-Marseille Universités & CNRS
Laboratoire Parole et Langage
pb@lpl.univ-aix.fr

Résumé. L'évolution récente des théories linguistiques conduit à une représentation différenciée des types d'information syntaxiques (par exemple la dominance et la précédence). Nous proposons dans cet article de systématiser cette approche pour conduire à une véritable représentation *décentralisée* de la syntaxe, nous permettant d'identifier un certain nombre de propriétés de base, devant être d'une façon ou d'une autre présente dans tout formalisme syntaxique.

Abstract. Modern linguistic theories tends to represent separately different kinds of syntactic information, (such as dominance or precedence). We propose in this paper to systematize this approach such as to obtain a *decentralized* representation, making it possible to identify a set of fundamental syntactic properties, to be represented in all syntactic formalism.

Mots-clés : Formalismes syntaxiques, propriétés syntaxiques, Model-Theoretic Syntax.

Keywords: Syntactic formalisms, basic syntactic properties, Model-Theoretic Syntax.

1 Introduction

Une des grandes idées proposées par GPSG ((Gazdar & al. 85)) consistait, dans le cadre des grammaires syntagmatiques, à représenter de façon distincte les informations hiérarchiques des informations de précédence linéaire. Derrière ce qui à première vue représente un intérêt en termes de représentation de l'information réside en fait une avancée majeure pour les théories linguistiques. Nous proposons dans cet article de pousser cette logique de différenciation de l'information syntaxique. Cette démarche nous permet de proposer une représentation beaucoup plus précise de ces informations et des interactions existant entre elles. Elle permet de plus d'ouvrir une alternative à une vision purement générative de la grammaire, sans en rejeter ce qui en fait l'intérêt (par exemple en termes de définition de classes d'objets). Cette vision, que nous appelons *décentralisée*, de l'information syntaxique repose sur la remise en question du statut central occupé dans tous les formalismes par la notion de dominance. Nous en faisons une information au même niveau que les autres, permettant ainsi la représentation de phénomènes complexes.

Cet article, plutôt que de présenter un formalisme particulier, précise les différents types d'information syntaxique qu'un formalisme doit représenter, si possible de façon explicite. Après

avoir situé les besoins auxquels doivent répondre les formalismes de haut niveau, nous proposons la description d'un ensemble de phénomènes illustrant les limites rencontrés par une représentation reposant prioritairement sur la dominance. Nous proposons donc une représentation totalement décentralisée, sous la forme d'un ensemble de propriétés devant être représentées en syntaxe.

2 La représentation de l'information syntaxique

2.1 L'évolution des théories

Le formalisme DI/PL propose d'expliciter le statut d'une information jusque là implicite : l'ordre linéaire[1]. Une représentation arborescente simple ne permet en effet pas de rendre compte de ce type d'information. Par ailleurs, GPSG propose de systématiser la représentation d'une partie des informations syntaxiques sous forme de traits associés aux catégories. Cette représentation de l'information permet l'introduction d'un nouveau type de relations comme les restrictions de cooccurrence de traits ou encore l'introduction de valeurs de traits par défaut. Il s'agit de contraintes portant directement sur la forme voire le contexte de réalisation d'une catégorie. Les informations concernant la structure hiérarchique sont ainsi complétées par d'autres types de contraintes permettant d'en contrôler la construction tout en précisant son contenu.

Les informations ainsi représentées sous forme de contraintes (cf. (Rogers97)) permettent de poser différemment la question du statut de la grammaire et sa relation par rapport au langage. Les approches génératives classiques proposent en effet de définir cette relation en termes de dérivation. La grammaire constitue ainsi un mécanisme d'énumération permettant de générer le langage. C'est ce que Pullum appelle la syntaxe générative énumérative (ou GES, cf. (Pullum & Scholz 01), (Huddleston & Pullum 02)). Une conception différente consiste à décrire la syntaxe dans la perspective de la théorie des modèles (on parle ainsi de *Model Theoretic Syntax*, ou MTS). Cette approche, présentée notamment dans (Blackburn & al. 93) puis développée dans (Backofen & al. 95), ou (Cornell & Rogers 00), consiste à concevoir l'analyse syntaxique comme une recherche de modèle dans un domaine spécifiée. Il s'agit d'une vision beaucoup plus souple du processus d'analyse, consistant, pour une affectation donnée (i.e. un ensemble de catégories) à vérifier les propriétés qu'elle satisfait. Une telle approche de l'analyse reposant sur les contraintes est présente de façon plus ou moins directe dans des théories comme HPSG (cf. (Sag al. 03)). Les théories basées sur les modèles n'ont donc pas pour objectif de construire une structure syntaxique en même temps qu'on vérifie la grammaticalité de l'input, mais plutôt de rechercher les propriétés qui sont vérifiées par cet énoncé. Les questions de la construction de la structure et de sa forme deviennent du même coup secondaire. En revanche, celles concernant l'identification des contraintes, et donc du type d'information à représenter, en plus de la dominance, devient centrale.

2.2 Les besoins

Les théories linguistiques, et donc les formalismes qui les représentent, doivent permettre rendre compte d'un certain nombre de phénomènes.

[1] Cette information n'est souvent prise en compte que de façon marginale, y compris dans les théories modernes (par exemple HPSG, cf. (Sag al. 03))

- **Non canonicité de l'entrée** : l'analyse de la langue parlée, mais également de matériel "tout venant", nécessite de rendre compte de phénomènes particuliers que l'on peut ranger globalement sous la rubrique "entrées non canoniques". Il s'agit d'entrées traditionnellement considérées comme mal formées pour différentes raisons : disfluences, structure incomplète, violation de règles, etc. Cet objectif est fondamental dans la plupart des approches modernes, il est même explicite dans certaines théories comme les grammaires de construction (cf. (Fillmore98)) ou l'optimalité (cf. (Prince93)).

- **Domaines** : l'information linguistique est répartie dans plusieurs domaines. Le traitement d'un message (sa compréhension) est le résultat de la convergence des parties d'information provenant des différents domaines (cf. (Blache02)). Il y a deux conséquences importantes à ce phénomène. Tout d'abord, les informations présentes dans chacun des domaines peuvent donc être partielles. Concrètement, chaque domaine n'est donc pas nécessairement porteur d'information pertinent pour la totalité de l'entrée. Pour ce qui concerne la syntaxe en particulier, cela signifie que toutes les unités identifiées n'entrent pas nécessairement dans une relation syntaxique. En d'autres termes, il n'est pas toujours possible (ni indispensable) de construire une structure syntaxique couvrant totalement l'entrée.

Par ailleurs, il est nécessaire de préciser comment les différents domaines interagissent. Par exemple, la vision compositionnelle classique de l'interaction syntaxe-sémantique consiste à calculer la représentation sémantique à partir de la structure syntaxique. De même, les relations prosodie/syntaxe sont généralement décrite en termes de superposition de structures (constituants prosodiques matchés aux constituants syntaxiques). Ces mécanisme posent donc problème dans le cas de structures partielles.

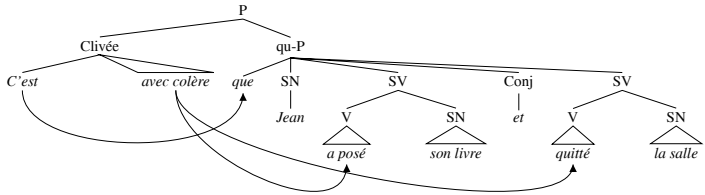
- **Motivation linguistique** : les informations portées dans un formalisme de haut niveau doivent être linguistiquement motivées et distinctes des éventuelles informations opérationnelles relevant de la mise en œuvre du formalisme plus que de la représentation de l'information linguistique. Par exemple, les mécanismes décrivant la propagation des traits ou contrôlant leur instanciation doivent être représentés séparément des propriétés linguistiques à proprement parler.

3 Les limites d'une représentation hiérarchique

Les approches syntagmatiques reposent toutes sur une hiérarchisation stricte de l'information. Le domaine décrit par les grammaires est donc celui des arbres. Les évolutions récentes, y compris celles remettant en cause les approches génératives strictes, consistent à adapter ce mode de représentation en le complétant de différentes manières et notamment en associant aux arbres des contraintes plus ou moins complexes, ce qui permet aux arbres d'avoir un degré de généricité plus ou moins grand : ensembles d'arbres locaux associés à des contraintes en DCG, ou à l'opposé schémas d'arbres soumis à des principes en HPSG. La notion d'arbre reste donc au cœur des représentations. Cela signifie (y compris pour les approches génériques comme HPSG) qu'il est nécessaire de construire un arbre pour en vérifier les contraintes. Les contraintes correspondant aux propriétés syntaxiques des objets analysés, cela signifie que la description d'une entrée repose sur la capacité de construire un arbre. Or, de nombreux exemples montrent les limites de ce type de représentation. Nous en relevons quelques uns.

- Les **dépendances à distance** : les constructions clivées, comme d'une façon plus générale les phénomènes d'extraposition, sont analysées en deux parties d'une même structure générale

comme décrit dans l'exemple suivant :

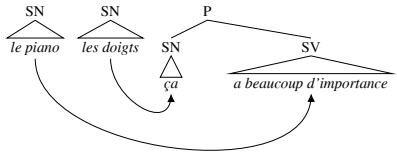


Cette représentation présente plusieurs lacunes, à commencer par la factorisation de l'auxiliaire dans la coordination, mais également concernant les relations entre l'élément clivé, son verbe recteur, ainsi que le pronom introduisant la *qu-phrase*. Ces relations sont selon les formalismes représentés sous la forme de traits se propageant à travers la structure (par exemple les traits *slash* en GPSG ou HPSG). Cela signifie que plusieurs types d'informations doivent être représentés : le découpage en unités, leur composition (informations représentées par l'arbre), mais également les dépendances entre ces unités (représentées dans la figure précédente par des flèches).

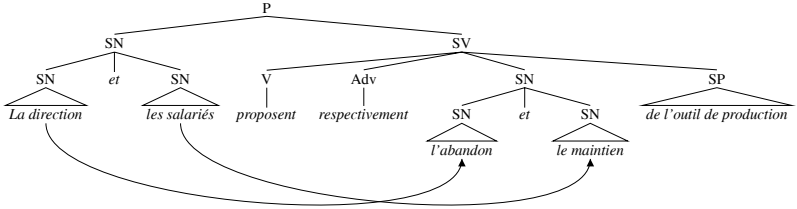
- **Extrapositions complexes** : plusieurs extrapositions peuvent se cumuler. L'exemple suivant[2] présente le cas d'une topicalisation doublé d'une dislocation :

(1) *le piano les doigts ça a beaucoup d'importance*

Il devient ici difficile (en tous cas peu pertinent) de proposer une structure arborescente unique. Il est en effet préférable de considérer qu'au niveau syntaxique, nous avons trois éléments dont les relations ne sont pas exprimées en termes de constituance, mais de dépendance.



- **Dépendances croisées** : ce type de phénomène intervient de façon systématique dans certaines langues. On le retrouve également souvent dans des constructions complexes telle la distribution sur des facteurs conjoints comme dans l'exemple suivant :



[2]Cet exemple, donné par José Deulofeu, est construit, mais censé relever de la langue parlée.

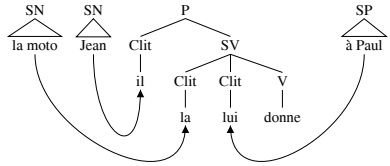
Cette construction illustre la nécessité d'indiquer des dépendances croisées construites par un verbe unique. Il est ici nécessaire de représenter à la fois les fonctions syntaxiques, en même temps que les relations existants entre les sujets et les objets.

- **Phénomènes paradigmatiques** : ces phénomènes fréquents à l'oral, mais également présents à l'écrit, consistent à réaliser plusieurs fois une même position de la structure syntaxique (cf. (Blanche-Benveniste84)). L'exemple suivant illustre à l'aide d'une grille un phénomène de ce type, appelé "entassement paradigmatique" (détaillé dans le cadre des Grammaires de Propriétés dans (Guénot06)) :

	les Anglais	qui ont quand même beaucoup d'humour	
euh	les journaux anglais		
enfin	les médias britanniques		ont fini par me répondre

Dans le cas, le noyau du SN sujet est réalisée plusieurs fois, il s'agit d'un phénomène de reprise dans lequel chaque nouvel élément remplace sémantiquement le précédent. Cet exemple présente en outre la possibilité dans ce type de construction de factoriser ce noyau qui sera modifié par la même relative (qui elle n'est réalisée qu'une fois). Il est clair qu'une structure arborescente n'est pas adaptée à la représentation de ce type de phénomène.

Un autre cas de ce type de phénomène concerne ce que (Blanche-Benveniste84) nomme le double marquage. Cela concerne typiquement les constructions disloquées dans lesquelles un élément référentiel est repris par un clitique :



En conclusion, les approches génératives classiques, en s'appuyant sur une représentation arborescente, accordent un statut particulier aux informations de dominance. Celles-ci doivent être vérifiées en préalable et les conditions de bonne formation sont exprimées en termes de contraintes sur les arbres. Il s'agit d'une limitation importante pour plusieurs raisons. Tout d'abord, nous avons vu qu'il n'est pas toujours possible ni souhaitable de représenter une structure syntaxique sous forme d'arbre. Par ailleurs, les exemples précédents montrent que les relations de dominance constituent une (petite) partie de l'information syntaxique. Celle-ci est complexe, composée d'informations différentes (que nous décrivons dans la section suivante) qui doivent être considérées comme également importantes, en tous cas de même niveau. Une telle conception décentralisée de l'information permet de rendre compte de constructions particulières du type de celles décrites ici. Notre proposition revient à dire qu'une structure n'est pas nécessairement *connexe* ou plus précisément, qu'elle n'est pas nécessairement connexe du point de vue de la dominance. En revanche, elle l'est si toutes les informations sont considérées comme des relations, au même titre et au même niveau que la dominance : les différentes unités de la structure décrivant un énoncé sont ainsi connectées par des relations variées.

4 Une représentation décentralisée

Plusieurs relations peuvent être proposées pour décrire l'information syntaxique. Nous cherchons ici à étendre la démarche initiée en GPSG consistant à représenter de façon séparée les différents types d'information. Cette section recense un ensemble qui nous paraît minimal d'informations qui doivent être représentées, quelque soit le formalisme. L'objectif est de rendre explicite ces informations comme GPSG a rendu explicites les relations d'ordre. Notre position est que chacune de ces propriétés est porteuse d'une partie de l'information, quelquefois de façon redondante. C'est l'interaction (ou la simple addition) de ces propriétés qui permet de fournir une description syntaxique précise d'un énoncé.

Les deux premières informations identifiées portent donc l'une sur la constituance, l'autre sur l'ordre.

- **Constituance** : une représentation en termes de constituants (correspondant donc dans une représentation arborescente à une information de dominance) est indispensable pour la définition des unités en termes syntagmatiques. Ce type de représentation reste aujourd'hui largement dominant dans les formalismes linguistiques. Elle permet tout d'abord d'exprimer des régularités en termes de substitution : des classes d'objets peuvent ainsi être décrites, partageant des propriétés communes. De plus, ce type de représentation permet également de donner des indications sur la fonction syntaxique en termes topologiques, par rapport à la position de l'élément relativement à d'autres.

- **Précédence linéaire** : GPSG a proposé de décrire le contrôle de l'ordre linéaire sous forme de déclarations. Il s'agit de contraintes contrôlant l'ordre entre éléments constituants d'une même unité, s'ils sont réalisés. Il est nécessaire de décrire de façon explicite ces paramètres. Certaines approches proposent de généraliser la précédence en l'étendant à des ensembles de catégories, permettant ainsi d'implanter des informations de type contextuel. Par ailleurs, il est utile d'étendre cette relation aux classes abstraites de catégories.

Nous trouvons également en GPSG d'autres types d'informations, qui ne sont pas représentées au même niveau que les deux précédentes. Il s'agit notamment de la restriction de cooccurrence de traits, permettant d'exclure la réalisation conjointe de certains types ou sous-types de catégories. Nous proposons de systématiser ce type d'information en représentant explicitement les relations de nécessité et d'impossibilité de cooccurrence.

- **Cooccurrence** : la relation de complémentation est typique de ce type de contrainte. La réalisation d'une catégorie particulière (souvent la tête) entraîne l'obligation de réaliser conjointement d'autres unités (les compléments). Le même type de relation existe également entre un spécifieur et la tête. Mais, au moins théoriquement, rien ne devrait empêcher la possibilité d'imposer la cooccurrence entre deux éléments, sans relation de rection, indépendamment de la tête.

(2) a. *Je me suis dit*.
b. *Je me le suis dit*.
c. **Je le suis dit*.

Il existe dans ce cas une contrainte de cooccurrence directement entre les compléments, sans intervention de la tête verbale.

- **Exclusion** : A l'inverse de la cooccurrence, la réalisation de certaines catégories peut entraîner l'impossibilité d'en réaliser d'autres. Comme pour la cooccurrence, ce type de contrainte

peut exister entre la tête et d'autres constituants (par exemple, le nom propre ne permet pas la construction d'un déterminant). Mais ici aussi ce type de relation peut exister en dehors de la tête. Ce cas est illustré dans l'exemple suivant :

(3) a. *Les trois livres.*
b. *Plusieurs livres.*
c. **Plusieurs trois livres.*

Il existe donc une contrainte particulière, indépendante de la tête nominale, entre le déterminant et l'adjectif. Celle-ci doit donc être représentée directement entre les constituants concernés.

- **Tête** : il est nécessaire dans une représentation syntagmatique d'identifier la tête d'une unité parmi l'ensemble de ses constituants[3]. Les informations de sous-catégorisation sont bien entendu importantes. Dans le type de représentation proposé ici, elles sont cependant portées par les contraintes de cooccurrence et instanciées au niveau du lexique. L'identification de la tête est cependant importante pour plusieurs raisons. Il s'agit en effet d'une catégorie obligatoirement réalisée dans une unité (éventuellement de façon isolée). Par ailleurs, elle permet de transmettre un certain nombre de caractéristiques à l'unité dont elle est la tête.

- **Unicité** : dans de nombreux cas, certaines catégories dans une unité ne peuvent pas être répétées. Cela concerne non seulement la tête (de façon systématique), mais également en fonction des unités d'autres constituants. Par exemple, un déterminant ne peut être répété dans un syntagme nominal. Il n'y a pas de régularité ni de prédictibilité particulière pour ce phénomène. Il est donc nécessaire de le représenter de façon explicite.

- **Dépendance** : Dans le grammaires de dépendance, cette relation est à la fois porteuse d'informations syntaxique et sémantique. Elle est de ce point de vue à la fois complémentaire et pour partie redondante avec certaines des informations exprimées précédemment. Nous proposons de représenter cette information, en particulier pour le rôle qu'elle joue dans la construction de la représentation sémantique, mais également car elle peut être porteuse de phénomènes de contrôle particuliers comme l'accord. Par ailleurs, rien n'impose que les relations de dépendances convergent toutes vers la tête. Plusieurs constructions, par exemple les effets de liste, illustrent ce type de phénomène :

(4) *J'en veux beaucoup beaucoup*

Dans cet exemple, la répétition de l'adverbe provoque un effet d'intensification. Les deux adverbes ne dépendent pas directement du verbe, nous préférons représenter le second adverbe comme dépendant du premier, cette relation ne passant pas par la tête verbale.

Ces différentes informations forment en quelque sorte le cœur de la structure syntaxique. Elles ne permettent cependant pas de décrire tous les phénomènes. Il est important de poursuivre l'examen de données variées de façon à exhiber de nouvelles propriétés devant être décrites de façon explicite. Nous en proposons deux nouvelles : l'une précisant l'ordre linéaire, la seconde exprimant des relations plus générales sur la distribution des unités.

- **Adjacence** : cette relation permet d'exprimer le fait que deux catégories doivent être juxtaposées. Ce type d'information est relativement peu fréquent et donc moins visible que les

[3]Certaines approches font jouer un rôle central à la tête (typiquement HPSG) en faisant transiter toutes les informations et tous les contrôles par cette catégorie. Cependant, même si d'un point de vue opérationnel ce choix peut être intéressant, rien ne l'impose.

autres contraintes. Pour autant, il est nécessaire de le représenter directement, plutôt que par une multiplication de contraintes de précédence. L'exemple suivant illustre ce phénomène :

(5) a. *C'est un homme aux doigts d'or.*

b. *C'est un homme tranquille aux doigts d'or.*

c. **C'est un homme aux doigts d'or tranquille.*

Dans ce cas, l'adjectif ne peut être séparé du nom par un *SP*, il doit lui doit être adjacent.

Les informations syntaxiques sont classiquement exprimées en termes d'unités et de relation entre ces unités. Il est cependant nécessaire d'ajouter une dimension supplémentaire permettant de définir des classes d'unités et d'exprimer des relations sur celles-ci. L'analyse de la langue parlée montre en effet la nécessité de dépasser le cadre de la simple distribution. Il convient en effet de décrire des phénomènes très fréquents, en particulier à l'oral, consistant à entasser des éléments sur un même paradigme de réalisation. Il est pour cela nécessaire d'identifier explicitement dans la structure syntaxique le paradigme de réalisation et d'en indiquer les contraintes.

- **Paradigme** : L'approche pronominale (voir (Blanche-Benveniste84)) a fourni une définition et une description de la notion de paradigmes. Elle repose sur la possibilité d'identifier des positions dans la structure syntaxiques en précisant la notion de substitution, lui substituant celle de proportionnalité (voir (van den Eynde & Mertens 03) pour une application au lexique). Cette relation permet d'identifier des classes d'arguments substituables (ou proportionnels) à un même pronom. Une structure prédicative est ainsi définie par la séquence des pronoms que le prédicat construit :

(6) a. *Je le leur dit.*

b. *Je leur en propose.*

Nous ne revenons pas ici sur la définition des paradigmes. Les trois premiers paradigmes (notés *P0, P1, P2*) correspondent aux fonctions sujet, objet direct, objet indirect en en précisant des sous-classes. Ce type de représentation d'information permet de préciser la structure argumentale des prédicats et donc de proposer un cadre pour la représentation des informations de sous-catégorisation. Celui-ci est ainsi décrit de façon très fine, en offrant également la possibilité de stipuler des relations de cooccurrence ou d'exclusion directement entre les pronoms identifiants des sous-classes.

Mais les paradigmes peuvent également jouer un rôle plus général. Il est en effet possible de proposer des contraintes telles que définies précédemment directement sur les paradigmes. Il est possible d'exprimer des contraintes générales sur l'unicité ou l'ordre des paradigmes, par exemple $P0 \prec P1 \prec P2$, cette contrainte pouvant être précisée voire remplacée par des contraintes locales spécifiques.

Par ailleurs, et il s'agit là d'une information essentielle, la connaissance du paradigme de réalisation permet de représenter de façon explicite ses réalisations pour un énoncé donné. Ainsi, dans le cas d'un entassement paradigmatique, il suffit d'indiquer l'appartenance de chaque unité de la liste à son paradigme, plusieurs unités pouvant apparaître dans le même paradigme. De même, les phénomènes de double marquage sont également représentés par l'indication des deux réalisations du même paradigme.

5 Représentation formelle

On peut préciser plus formellement les propriétés décrites précédemment. On utilise dans la suite les notations suivantes : x, y (minuscules) pour représenter les variables individuelles ; X, Y (majuscules) les variables ensemble. On note $C(x)$ l'ensemble des variables individuelles dans le domaine de la catégorie C (voir (Backofen & al. 95) pour plus de précisions). On utilise les prédicats binaires représentant la constituance ($\triangleleft$), la précédence linéaire ($\prec$) et l'égalité ($\approx$). Dans ce qui suit, on considère que le domaine est celui des graphes, plutôt que des arbres. Les variables individuelles représentent des nœuds de ces graphes. La définition des informations de constituance consiste en une déclaration d'ensemble, de même que la notion de paradigme. Les définitions suivantes concernent les autres informations.

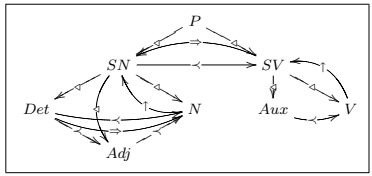
- *Unicité(A)* : $(\forall x, y)[A(x) \wedge A(y) \rightarrow x \approx y]$
  Si un élément de catégorie A est réalisé, il ne peut exister d'autre nœud de même catégorie. L'unicité précise donc les constitutants ne pouvant être répétés dans une construction.
- *Précédence(A,B)* : $(\forall x, y)[(A(x) \wedge B(y) \rightarrow y \not\prec x)]$
  Cette définition indique que si les nœuds x et y sont réalisés, alors y ne peut pas précéder x.
- *Tête(A)* : $(\exists x)(\forall y)[A(x) \wedge A(y) \rightarrow x \approx y]$
  Il existe un nœud x de catégorie A et il n'existe pas d'autre nœud y de catégorie indentique. Une tête n'est réalisée qu'une fois dans une construction.
- $Cooccurrence(A, B) : (\forall x, y)[A(x) \rightarrow B(y)]$
  Si un nœud x de catégorie A est réalisé, un nœud y de catégorie B doit l'être également.
- $Exclusion(A, B) : (\forall x)(\nexists y)[A(x) \wedge B(y)]$
  S'il existe un nœud x, il ne peut exister un nœud de même niveau y.
- $Adjacence(A, B) : (\forall C)(\forall x, y, z)[A(x) \wedge B(y) \wedge C(z) \wedge z \triangleleft x \wedge z \triangleleft y \rightarrow (\forall t)[y \prec t \vee t \prec x]]$
  Si deux nœud x et y sont réalisés comme constituants d'une même catégorie C, alors tout autre nœud t devra soit les précéder, soit les suivre.

Au total, on peut donc définir une grammaire comme un n-uplet de la forme :

$$G = \langle W, \Rightarrow, \otimes, \circ, \triangleleft, \uparrow, \prec, \theta, \mathcal{P}, \frown \rangle$$

dans lequel les symboles représentent les relations suivantes : $\Rightarrow$ (cooccurrence), $\otimes$ (exclusion), $\circ$ (unicité), $\triangleleft$ (constituance), $\uparrow$ (tête), $\prec$ (précédence), $\frown$ (adjacence) et $\mathcal{P}$ (paradigmes).

La figure suivante illustre un ensemble de relations pouvant être spécifiées entre différentes unités. Chaque relation est typée par le symbole de la propriété qu'elle porte.



6 Conclusion

Les formalismes syntaxiques de haut niveau doivent permettre la représentation de l'information syntaxique sans dépendre de contrainte opérationnelles spécifiques. En particulier, l'évolution récente des théories linguistique a montré la nécessité d'une part de distinguer les différents types d'information et d'autre part de les traiter au même niveau. Il devient alors possible de se libérer de la nécessité de construire une structure syntaxique sous la forme d'un arbre avant de pouvoir vérifier les autres propriétés décrivant l'énoncé analysé. La décentralisation de l'information permet de plus d'identifier explicitement l'ensemble des informations syntaxiques qui entrent en jeu dans une description. Nous avons proposé dans cet article un ensemble de relations décrivant les informations syntaxiques. Ces relations ont vocation à faire partie de tout formalisme de haut niveau.

Références

BLACKBURN P., GARDENT C. & MEYER-VIOL W. (1993). Talking About Trees. In *Proceedings of EACL*.

BACKOFEN R., ROGERS J. & VIJAY-SHANKER K. (1995). A First-Order Axiomatization of the Theory of Finite Trees. *Journal of Logic, Language, and Information*, 4 :1.

BLACHE P. & DI CRISTO A. (2002). Variabilité et dépendances des composants linguistiques. In *actes de TALN-2002*.

BLACHE P. (2005). Property Grammars : A Fully Constraint-Based Theory. In H. CHRISTIANSEN & al. (eds), *Constraint Solving and NLP*, Lecture Notes in Computer Science, Springer.

BLANCHE-BENVENISTE C., DEULOFEU J., STEFANINI J. & VAN DEN EYNDE K. (1984). *Pronom et syntaxe. L'approche pronominale et son application au français*, Selaf.

CORNELL T. & ROGERS J. (2000). Model Theoretic Syntax. In L. LAI-SHEN CHENG & R. SYBESMA (eds), *The Glot International State of the Article Book I*, Holland Academic Graphics.

FILLMORE C. (1998). Inversion and Contructional Inheritance, in *Lexical and Constructional Aspects of Linguistic Explanation*, Stanford University.

GAZDAR G., KLEIN E., PULLUM G. & SAG I. (1985). *Generalized Phrase Structure Grammars*, Blackwell.

GUÉNOT M.-L. (2006). *Eléments de grammaire du français : pour une théorie descriptive et formelle de la langue*, Thèse de doctorat, Université de Provence.

HUDDLESTON R. & PULLUM G. (2002). *The Cambridge Grammar of the English Language*, Cambridge University Press.

PRINCE A. & SMOLENSKY P. (1993). In *Optimality Theory : Constraint Interaction in Generative Grammars*, Technical Report RUCCS TR-2, Rutgers Center for Cognitive Science.

PULLUM G. & SCHOLZ B. (2001). On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In *Proceedings of the conference on Logical Aspects of Computational Linguistics*, Springer.

ROGERS J. (1997). Grammarless Phrase Structure Grammar. *Linguistics and Philosophy*, 20.

SAG I., WASOW T. & BENDER E. (2003). *Syntactic Theory. A Formal Introduction*, CSLI.

VAN DEN EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13.

Une grammaire d'interaction du français

Guy PERRIER
LORIA - Université Nancy 2
perrier@loria.fr

Résumé. Nous présentons une grammaire du français à relativement large couverture dans le formalisme des grammaires d'interaction. Ce formalisme combine deux idées-forces : la grammaire est vue comme un système de contraintes à travers la notion de description d'arbre, et la sensibilité aux ressources de la langue est utilisée comme principe de composition syntaxique à l'aide de la notion de polarité. Nous donnons un aperçu du pouvoir expressif du formalisme en modélisant quelques phénomènes linguistiques significatifs et nous montrons que l'architecture de la grammaire répond à un souci de réutilisabilité et de faisabilité, crucial quand on cherche à construire des ressources à large couverture : distinction entre une grammaire source modulaire et une grammaire objet obtenue par compilation de la première, indépendance du lexique par rapport à la grammaire. Enfin, nous présentons les résultats d'une évaluation de la grammaire sur une suite de phrases tests, effectuée à l'aide de l'analyseur syntaxique LEOPAR.

Abstract. We present a French grammar with a relatively large coverage in the formalism of Interaction Grammars. This formalism combines two key ideas : the grammar is viewed as a constraint system, which is expressed through the notion of tree description, and the resource sensitivity of the language is used as a syntactic composition principle by means of the notion of polarity. We give an outline of the expressivity of the formalim by modelling significative linguistic phenomena and we show that the grammar architecture provides for reusability and tractability, which is crucial for building large coverage resources : a modular source grammar is distinguished from the object grammar which results from the compilation of the first one, the lexicon is independent of the grammar. Finally, we present the results of an evaluation of the grammar with a test suite of sentences achieved with the LEOPAR parser.

Mots-clés : syntaxe, grammaire formelle, méta-grammaire, grammaire d'interaction.

Keywords: syntax, formal grammar, meta-grammar, interaction grammar.

1 Introduction

Le travail que nous présentons ici s'inscrit dans une démarche de modélisation des langues à partir de connaissances linguistiques qui fait une place centrale à l'expérimentation. Dans cet objectif, il est nécessaire d'exprimer ces connaissances linguistiques sous forme de grammaires et de lexiques avec la couverture la plus large possible tant en termes de phénomènes linguistiques représentés que de mots auxquels ils s'appliquent. Or, on sait combien il est difficile de construire de telles ressources.

La première difficulté est celle du choix du formalisme pour représenter la grammaire. Ac-

tuellement, il n'y a pas vraiment de formalisme qui prévaut dans la communauté scientifique. Ceux qui sont les plus répandus ont tous leurs points forts et leurs points faibles. Si nous avons conçu un nouveau formalisme, celui des Grammaires d'Interaction (GI), c'est pour faire la synthèse de deux idées importantes exprimées jusqu'ici dans deux types de formalismes différents : l'utilisation de la sensibilité aux ressources des langues comme principe de composition syntaxique qui est un trait caractéristique des grammaires catégorielles (Retoré, 2000) et la vision des grammaires comme systèmes de contraintes qui est celle des grammaires d'unification telles que LFG (Bresnan, 2001) ou HPSG (Sag *et al.*, 2003).

Même si nous utilisons un formalisme original, notre souci est celui de la réutilisabilité, souci qui s'exprime de deux façons :
– Comme pour la conception des langages de programmation, nous distinguons deux niveaux dans la grammaire : la *grammaire source*, qui est écrite par l'humain (le linguiste dans l'idéal) et qui permet d'exprimer les généralisations linguistiques, et la *grammaire objet* qui est directement utilisable par un système de TAL. La première est compilée dans la seconde et nous avons utilisé pour cela XMG (Duchier *et al.*, 2005). XMG fournit un langage de haut niveau pour écrire une grammaire source et un compilateur qui traduit cette grammaire source en une grammaire objet opérationnelle.
– La grammaire est aussi conçue de telle façon qu'elle puisse s'interfacer avec un lexique indépendant du formalisme où les entrées se présentent comme des structures de traits.

C'est de cette manière que nous avons construit une grammaire du français à relativement large couverture dans le formalisme des GI et le but de l'article est de présenter cette grammaire.

2 Les grammaires d'interaction

Les GI[1](Perrier, 2004) sont un formalisme grammatical dédié à la syntaxe et à la sémantique des langues naturelles qui s'appuie sur deux notions, celle de *description d'arbre* et celle de *polarité*.

2.1 Les descriptions d'arbres

Dans une vision dérivationnelle de la syntaxe des langues, les objets syntaxiques manipulés sont en général des arbres qui sont composés de façon plus ou moins sophistiquée (grammaires algébriques, grammaires d'arbres adjoints, grammaires catégorielles . . .). Empruntant notre vision à la théorie des modèles (Pullum & Scholz, 2001), nous ne manipulons pas directement des arbres syntaxiques mais des propriétés permettant de les décrire, autrement dit des descriptions d'arbres (Rogers & Vijay-Shanker, 1994). Cette approche est très souple en ce sens qu'elle permet d'exprimer de façon totalement indépendante des propriétés élémentaires d'arbres que l'on peut ensuite combiner librement.

Une description d'arbre peut être vue, soit comme un arbre sous-spécifié, soit comme une spécification d'une famille d'arbres, chaque arbre étant un modèle de cette spécification. La figure 1 donne un exemple de description d'arbre associée au pronom relatif *qui*, lorsqu'il est employé comme complément indirect. Cet emploi donne lieu au phénomène complexe d'une double dépendance non bornée (le *pied piping* en anglais) comme l'illustrent les exemples suivants qui

[1]Pour une présentation complète des GI, le lecteur pourra se reporter à l'article (Perrier, 2004).

sont tous couverts par la description de la figure 1[2].

(a) *Jean [à* **qui***] Pierre a présenté Marie* □ *est ingénieur.*
(b) *Jean [à la femme de* **qui***] Pierre a présenté Marie* □ *est ingénieur.*
(c) *Jean [à la femme de* **qui***] Pierre sait qu'on a présenté Marie* □ *est ingénieur.*

Une description est un ensemble fini de nœuds structurés par deux types de relations : *domination* et *précédence*. Les nœuds, qui représentent des syntagmes, sont étiquetés par des traits décrivant leurs propriétés morpho-syntaxiques. Les valeurs des traits sont des atomes ou des disjonctions d'atomes et elles peuvent être partagées grâce à un mécanisme de co-indexation[3]. Les nœuds peuvent être typés : ils peuvent porter la propriété *Empty* (en fond blanc sur la figure 1) ou *Full*, selon qu'ils ont une forme phonologique vide ou pleine ; ils peuvent porter la propriété *Anchor* (cadre double sur la figure 1), s'ils représentent un nœud ancrant un mot de la langue.

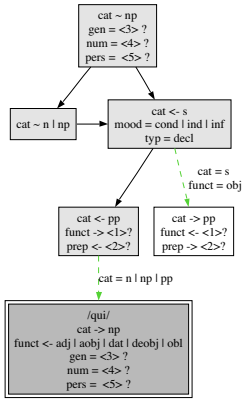


FIG. 1 – Description d'arbre associée au pronom relatif *qui* utilisé dans un complément indirect

Voici les relations entre nœuds qui sont utilisées pour définir les descriptions d'arbres :

1. *Relations de domination* :
   - $A \rightarrow B$ signifie que A est le père de B (flèche vers le bas continue sur la figure 1).
   - $A \rightarrow [B$ signifie que A est le père de B et qu'il n'a pas d'autre fils qui précède B.
   - $A \rightarrow B]$ signifie que A est le père de B et que B ne précède aucun autre fils de A.
   - $A \rightarrow * B$ signifie que A domine largement B (clôture réflexive et transitive de la première relation représentée par une flèche discontinue vers le bas sur la figure 1).
   - $A \rightarrow * [t_1 = v1, \ldots, t_n = v_n] B$ signifie en plus que tout nœud strictement dominé par A et dominant strictement B doit être étiqueté par une structure de traits subsumée par la contrainte $[t_1 = v_1, \ldots, t_n = v_n]$[4].

[2]Le groupe prépositionnel extrait est placé entre crochets et sa trace dans la proposition relative est représentée par le symbole □.

[3]Lorsque deux traits partagent une même valeur, un indice commun <n> est placé devant leurs valeurs. Lorsqu'un trait a comme valeur la disjonction de tous les atomes de son domaine, cette valeur est notée " ?".

[4]Dans l'implémentation actuelle de la grammaire, le sens de la contrainte est un peu différent dans la mesure où elle s'applique aussi aux deux nœuds reliés par la domination large.

2. *Relations de précédence* :
   - $A >> B$ signifie que A précède immédiatement B (flèche horizontale continue sur la figure 1).
   - $A >* B$ signifie que A précède B (clôture transitive de la précédente relation représentée graphiquement par une flèche horizontale discontinue).

2.2 Les polarités

Les polarités permettent d'exprimer l'état de saturation des arbres syntaxiques. Attachées à des traits qui décorent les nœuds des descriptions, elles ont la signification suivante :
- un trait positif $t \rightarrow v$ exprime une ressource disponible qui doit être consommée ;
- un trait négatif $t \leftarrow v$ exprime une ressource attendue qui doit être fournie ; c'est le dual d'un trait positif ;
- un trait neutre $t = v$ exprime une propriété linguistique qui ne se comporte pas comme une ressource consommable.
- un trait virtuel $t \sim v$ exprime une propriété qui a besoin de se réaliser en se combinant avec un trait réel (positif, négatif ou neutre).

Sur la figure 1, le nœud vide représentant la trace du syntagme prépositionnel extrait de la proposition relative est porteur d'un trait positif $cat \rightarrow pp$ et d'un trait négatif $funct \leftarrow \langle 1 \rangle ?$, qui signifie que ce nœud fournit un groupe prépositionnel qui attend de recevoir une fonction syntaxique. La racine de l'arbre porte un trait virtuel $cat \sim np$ qui signifie que le nœud représente un syntagme nominal virtuel qui doit se combiner avec un syntagme nominal réel.

Les descriptions décorées par des structures de traits polarisés prennent alors la forme de *descriptions d'arbres polarisées (DAP)*.

2.3 La grammaire comme système de contraintes

Une grammaire d'interaction particulière est définie par un ensemble fini de DAP élémentaires qui engendre un langage d'arbres. Un arbre du langage est défini comme un arbre syntaxique modèle d'un ensemble fini d'arbres élémentaires de la grammaire vérifiant deux propriétés particulières : il est à la fois *saturé* et *minimal*.
- Saturé, il réalise une neutralisation complète des polarités présentes ; chaque trait positif $t \rightarrow v$ doit rencontrer dans le modèle son dual $t \leftarrow v$ et vice-versa ; chaque trait virtuel doit rencontrer dans le modèle un trait réel correspondant.
- Minimal, le modèle doit ajouter un minimum d'information à celle présente dans les descriptions initiales (il ne peut ajouter ni relation de domination immédiate, ni trait qui ne sont pas présents dans les descriptions de départ).

L'analyse syntaxique se ramène alors à la résolution d'un système de contraintes. Elle consiste à construire tous les modèles saturés et minimaux d'un ensemble fini de DAP élémentaires. Dans la pratique, la grammaire que nous avons construite est totalement lexicalisée : toute DAP élémentaire possède une ancre unique qui lui permet de se lier à un mot de la langue. La lexicalisation permet, pour l'analyse d'une phrase, de sélectionner uniquement des DAP ancrées par des mots de la phrase. Une fois l'ensemble des DAP sélectionné, la construction d'un modèle saturé et minimal se fait pas à pas à l'aide d'une opération de fusion de nœuds deux par deux, guidée par l'une ou l'autre des contraintes suivantes :

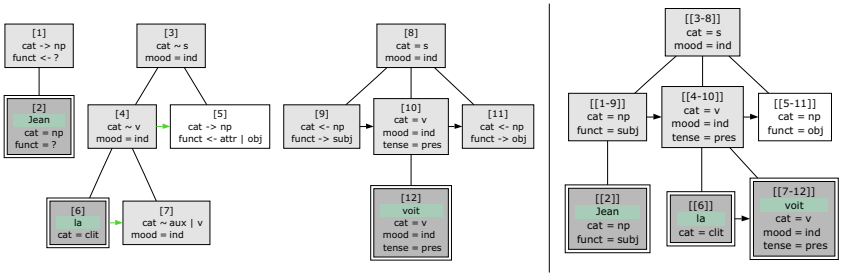


FIG. 2 – DAP associée à la phrase *Jean la voit* et son modèle saturé minimal

– neutraliser un trait positif avec un trait négatif de même nom et porteur d'une valeur qui s'unifie avec celle du premier trait ;
– réaliser un trait virtuel à l'aide d'un trait réel (positif, négatif ou neutre) de même nom et porteur d'une valeur qui s'unifie avec celle du premier trait.

Les contraintes induites par la description font que la fusion de deux nœuds entraîne généralement une superposition partielle de leurs contextes représentés par les fragments d'arbres dans lesquels ils se situent.

Ainsi, les GI combinent les points forts de deux familles de formalismes : la souplesse des *grammaires d'unification* et le contrôle de saturation des *grammaires catégorielles*.

La figure 2 présente un exemple d'analyse syntaxique, celle de la phrase *Jean la voit*[5]. La partie gauche montre l'ensemble des DAP initiales associées par la grammaire à la phrase. La grammaire étant lexicalisée, chacune des DAP est associée à un mot de la phrase et a été extraite d'un lexique. Ces DAP ont été réunies en une seule à laquelle on a ajouté une information de précédence entre les ancres, qui n'apparaît pas sur le schéma, pour prendre en compte l'ordre des mots de la phrase.

Le passage de la description initiale à son modèle donné par la partie droite de la figure 2 est réalisé par une suite de 3 fusions de nœuds[6]. Le simple jeu des contraintes d'arbre fait que ces 3 fusions en entraînent deux autres ainsi qu'une superposition partielle d'arbres.

3 Le pouvoir d'expression des grammaires d'interaction

Dans les limites de cet article, nous ne pouvons étudier de façon exhaustive cette question et nous avons choisi d'en illustrer trois aspects particulièrement significatifs :
– les relations de domination large avec contraintes pour représenter les dépendances non bornées en cascade (pied piping),
– les polarités positives et négatives pour modéliser les paires de mots exprimant la négation,
– les polarités virtuelles pour exprimer la position relativement libre des modificateurs de phrases.

[5]Nous avons simplifié la figure en ne mentionnant pas les traits d'accord.

[6]Dans l'entête de chaque nœud du modèle, on peut retrouver le numéro des nœuds des DAP initiales qui ont été fusionnées.

3.1 Dépendances non bornées et relations de domination larges

Les relations de domination large sont utilisées pour représenter les dépendances non bornées et les structures de traits qu'il est possible d'associer à ces relations permettent d'exprimer des contraintes sur ces dépendances, par exemple les barrières à l'extraction.

Les pronoms relatifs, tels que *qui* ou *lequel*, donnent lieu à des dépendances non bornées en cascade (pied piping) comme dans la phrase : *Jean [dans l'entreprise de* **qui***] Marie sait que l'ingénieur travaille* □ *est malade* :

– Il y a une première dépendance non bornée entre le verbe *travaille* et son complément extrait *dans l'entreprise de qui*. La trace du complément extrait est marquée par le symbole □. Cette dépendance est modélisée dans la DAP associée au pronom relatif *qui* représentée sur la figure 1 par une relation de domination large. La contrainte associée à cette relation de domination exprime que la dépendance du syntagme prépositionnel par rapport au verbe dont il est complément ne peut traverser qu'une suite indéterminée de complétives ou d'infinitives objet. Cela permet de refuser la phrase suivante : ∗ *Jean [dans l'entreprise de* **qui***] Marie qui travaille* □ *le connaît est malade*.

– A l'intérieur du syntagme prépositionnel, il y a une deuxième dépendance non bornée entre la tête du syntagme et le pronom relatif *qui*, qui peut être enchâssée plus ou moins profondément dans ce syntagme. Cette dépendance est aussi représentée sur la figure 1 par une relation de domination large et la contrainte associée exprime que les syntagmes enchâssés sont des noms communs, des syntagmes nominaux ou prépositionnels. Cela permet de refuser la phrase : ∗ *Jean [dans l'entreprise qui appartient à* **qui***] Marie travaille* □ *est malade*.

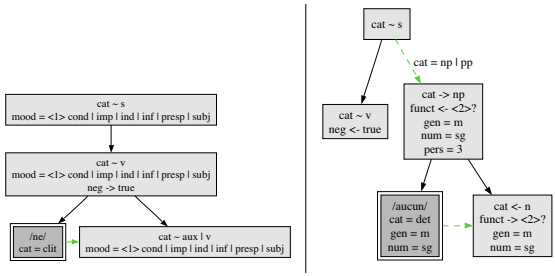


FIG. 3 – DAP associées respectivement à la particule *ne* et au déterminant *aucun*

3.2 L'utilisation des polarités pour modéliser la négation

En français, la négation peut s'exprimer à l'aide de la particule *ne* couplée avec un mot qui peut être un déterminant, un pronom ou un adverbe. La position de la particule *ne* est figée avant un verbe porteur d'une inflexion mais l'autre mot, s'il s'agit d'un déterminant comme *aucun* ou un pronom comme *personne*, peut avoir une position relativement libre dans la phrase, comme le montrent les exemples suivants :

(a) *Jean* **ne** *parle à* **aucun** *collègue.*
(b) *Jean* **ne** *parle à la femme d'***aucun** *collègue.*
(c) **Aucun** *collègue de Jean* **ne** *parle à sa femme.*

Comme le montre la figure 3, le couplage de *ne* avec *aucun* est exprimé par un trait polarisé *neg*

porté par le nœud représentant la projection maximum du noyau verbal : *aucun* est en attente d'un tel trait qui va être fourni par *ne*. La position relativement libre de *aucun* est exprimée par une domination large du nœud représentant la proposition sur le syntagme nominal qu'il introduit. La contrainte associée à cette domination large exprime le fait que *aucun* ne peut introduire que des arguments du verbe tête de la phrase ou leurs compléments. Bien entendu, tous les usages ne sont pas couverts par ces deux descriptions et il est notamment nécessaire de modifier légèrement celle associée à *aucun* pour analyser une phrase comme *Jean* **ne** *voit* **jamais aucun** *responsable*. Il faut ajouter une nouvelle entrée pour *aucun* avec un trait *neg* virtuel au lieu d'être négatif.

3.3 L'adjonction de modificateurs à l'aide de polarités virtuelles

En français, la place des compléments circonstanciels dans la phrase est relativement libre, comme le montrent les exemples suivants :

(a) **Le soir**, *Jean va rendre visite à Marie.*
(b) *Jean,* **le soir**, *va rendre visite à Marie.*
(c) *Jean va rendre visite* **le soir** *à Marie.*
(d) *Jean va rendre visite à Marie* **le soir**.

Ces variantes expriment des intentions communicatives différentes mais *le soir* est dans tous les cas un complément circonstanciel, modificateur de la phrase.

La polarité *virtuelle* $f \sim v$ n'existait pas dans la version précédente des GI (Perrier, 2004). L'adjonction de modificateurs était effectuée comme dans beaucoup de formalismes (grammaires d'arbres adjoints, grammaires catégorielles ...) par ajout d'un niveau supplémentaire dans l'arbre syntaxique où était le syntagme modifié : à la place d'un nœud de catégorie X était inséré un arbre formé d'une racine de catégorie X et de ses deux fils : le modificateur et le syntagme initial de catégorie X objet de la modification. Si cette introduction d'un niveau supplémentaire est parfois justifiée, le plus souvent elle vient introduire une complexité et une ambiguïté artificielles. Reprenant une idée de (Nasr, 1995) avec son système de polarités noires et blanches, nous avons introduit les polarités virtuelles. Cela nous permet d'ajouter un modificateur comme fils supplémentaire du nœud qu'il modifie sans rien changer au reste de l'arbre syntaxique dans lequel se situe le nœud modifié. La DAP de la figure 1 en donne un exemple car la proposition relative représentée par le fils droit de la racine y apparaît comme un modificateur de groupe nominal. En anglais, on parle alors de *sister adjunction* et elle est utilisée dans certains formalismes (grammaires de dépendance, grammaires de substitution de descriptions (Rambow *et al.*, 2001)). Cette modélisation des modificateurs est beaucoup plus souple que la précédente et nous a permis de traiter les exemples présentés ci-dessus sans difficulté, ainsi que les propositions incises et incidentes, considérées comme des modificateurs de phrases.

4 L'architecture de la grammaire

4.1 L'organisation modulaire de la grammaire

La grammaire a été construite avec l'outil XMG (Duchier *et al.*, 2005) qui permet d'écrire des grammaires d'un haut niveau d'abstraction sous une forme modulaire et de les compiler

ensuite dans des grammaires de plus bas niveau utilisables par des systèmes de TAL. Décrivons brièvement les traits caractéristiques de XMG.

Une grammaire est organisée en une hiérarchie de classes à l'aide de deux opérations de composition : *conjonction* et *disjonction*. Elle est aussi structurée selon plusieurs dimensions qui se retrouvent dans chaque classe. Notre grammaire n'utilise que deux dimensions : la première est la dimension syntaxique où les objets sont des DAP et la seconde est celle de l'interface avec le lexique où les objets sont des structures de traits.

Pour définir la conjonction de deux classes, il est nécessaire de préciser la manière dont les composantes de chaque dimension se combinent :
– pour la dimension syntaxique, c'est l'union des DAP qui est effectuée ;
– pour la dimension des interfaces avec le lexique, c'est l'unification entre structures de traits qui est réalisée.
Pour éviter les collisions d'identificateurs, leur portée est strictement contrôlée ; par défaut, elle est locale à la classe où ils sont déclarés, mais l'on peut exporter un identificateur qui devient alors visible pour l'extérieur. Lorsque l'on combine deux classes, leurs identificateurs exportés doivent être disjoints ; si l'on veut confondre deux identificateurs, il faut le dire explicitement à l'aide d'une équation. A la différence de (Crabbé, 2005) qui utilise un système de couleurs, nous avons choisi de nous servir d'un nombre extrêmement limité de noms de nœuds, pertinents linguistiquement, pour contrôler la conjonction des classes. XMG ne permettant pas de définir des identificateurs globaux, nous avons utilisé des équations entre identificateurs pour contourner le problème.

La grammaire actuelle comprend 448 classes dont 121 classes terminales, qui sont compilées en 2059 DAP. Ces classes sont rangées par famille. Une famille peut être réutilisée par les classes d'une autre. C'est le cas par exemple de la famille *Complement* qui contient les classes définissant les compléments de structures prédicatives. Elle est utilisée par 3 autres familles : *Adjective*, *Noun* et *VerbDiathesis*, qui décrivent respectivement le comportement syntaxique des adjectifs, celui des noms et les différentes diathèses du verbe. La famille *VerbDiathesis* utilise aussi les familles *Verbmorphology* et *Verbfunction* qui décrivent respectivement la morphologie verbale et les fonctions que peut occuper le verbe dans la phrase (y compris lorsqu'il est participe présent et participe passé).

4.2 La liaison avec un lexique indépendant du formalisme

La grammaire, dans sa forme actuelle, est totalement lexicalisée : chaque DAP élémentaire de la grammaire a un unique nœud ancre destiné à être associé à un mot de la langue. Pour cela, chaque DAP est associée à une structure de traits qui décrit de façon indépendante du formalisme un cadre syntaxique correspondant aux mots pouvant ancrer la description. Cette structure de traits constitue l'*interface* de la DAP avec le lexique.

L'ensemble des traits utilisés dans les interfaces sont différents de ceux utilisés dans les descriptions car leur rôle n'est pas le même : ils visent non pas à décrire des syntagmes mais les mots de la langue et ceci d'une façon indépendante du formalisme.

La figure 4 présente dans sa partie gauche une DAP non ancrée correspondant à un verbe transitif à un temps fini dans une configuration canonique. Cette description est accompagnée de son interface et on y a fait figurer les indices associés à certaines valeurs qui montrent que certains traits de l'interface partagent leur valeur avec des traits de la description.

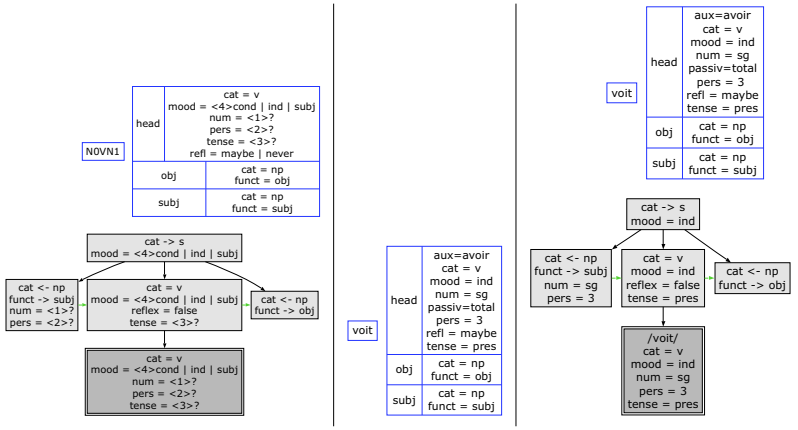


FIG. 4 – DAP non ancrée associée à un verbe transitif, entrée lexicale pour le verbe *voit* et DAP ancrée par le verbe *voit*

Le lexique des mots de la langue associe ceux-ci à des cadres syntaxiques dans un format identique aux interfaces des descriptions. La figure 4 , dans sa partie centrale, montre une entrée d'un tel lexique pour le verbe *voit* : il décrit le cadre syntaxique associé à *voit* lorsque celui-ci est employé comme verbe transitif.

L'ancrage des descriptions de la grammaire se fait ensuite par unification de leurs interfaces avec les cadres syntaxiques compatibles du lexique. Un mécanisme de co-indexation entre valeurs de traits de la description et de l'interface permet un paramétrage de certains traits. La figure 4 montre dans sa partie droite la DAP obtenue par ancrage de la DAP de gauche à l'aide de l'entrée lexicale présentée au centre. Cet ancrage a consisté à unifier l'interface de la DAP non ancrée avec le cadre syntaxique offert par le lexique.

5 Evaluation sur une suite de phrases tests

Notre but est d'évaluer le plus finement et de la façon la moins coûteuse possible la couverture de notre grammaire. Une réponse adaptée est d'utiliser une suite de phrases tests mais il est important que cette suite contienne non seulement des exemples positifs mais aussi des exemples négatifs pour évaluer le pouvoir de surgénération de la grammaire.

Nous avons choisi l'une des rares suites de ce type qui existe pour le français : la TSNLP (Lehmann *et al.*, 1996) qui comprend 1690 phrases positives et 1935 phrases négatives. Elle est loin de couvrir toute la grammaire du français ; notamment, elle contient très peu de phrases complexes et par contre, elle s'attarde beaucoup sur certains phénomènes tels que la coordination ou l'ordre des compléments circonstanciels dans la phrase. On peut même dire que notre grammaire prend en compte des phénomènes ignorés de la TSNLP : voies passive et moyenne, sous-catégorisation des noms et adjectifs prédicatifs, contrôle du sujet des infinitives compléments, propositions relatives, interrogatives . . .

Pour effectuer l'analyse, nous avons utilisé LEOPAR[7], qui est un analyseur syntaxique fondé sur les GI. Avec la grammaire actuelle, il accepte 88% des 1690 phrases positives et rejette 85% des 1935 phrases négatives de la TSNLP. Les 15% de phrases négatives acceptées le sont essentiellement par absence d'intégration de règles phonologiques et de la sémantique dans la grammaire. Les 12% des phrases positives non couvertes le sont pour des raisons très diverses : phrases du langage parlé prenant certaines libertés avec la grammaire, expressions figées ou semi-figées, phénomènes non encore pris en compte (constructions causatives, superlatifs . . .).

6 Perspectives

D'ores et déjà, il est possible d'utiliser LEOPAR avec un lexique à large couverture pour analyser des corpus tout venant. Il est nécessaire d'enrichir la grammaire pour couvrir un certain nombre de phénomènes linguistiques courants non encore pris en compte. Il faudra aussi améliorer les performances de l'analyseur pour faire face à l'explosion potentielle résultant de l'augmentation de la taille de la grammaire se conjuguant avec la longueur des phrases.

Références

BRESNAN J. (2001). *Lexical-Functional Syntax*. Oxford : Blackwell Publishers.

CRABBÉ B. (2005). *Représentation informatique de grammaires fortement lexicalisées : application à la grammaire d'arbres adjoints*. thèse de doctorat, université Nancy2.

DUCHIER D., LE ROUX J. & PARMENTIER Y. (2005). XMG : Un compilateur de métagrammaires extensible. In *TALN 2005, Dourdan, France*.

LEHMANN S., OEPEN S., REGNIER-PROST S., NETTER K., LUX V., KLEIN J., FALKEDAL K., FOUVRY F., ESTIVAL D., DAUPHIN E., COMPAGNION H., BAUR J., BALKAN L. & ARNOLD D. (1996). TSNLP — Test Suites for Natural Language Processing. In *Proceedings of COLING 1996, Kopenhagen*.

NASR A. (1995). A formalism and a parser for lexicalised dependency grammars. In *4th International Workshop on Parsing Technologies (IWPT)*.

PERRIER G. (2004). La sémantique dans les grammaires d'interaction. *Traitement Automatique des Langues*, **45**(3), 123–144.

PULLUM G. K. & SCHOLZ B. C. (2001). On the Distinction between Model-Theoretic and Generative-Enumerative Syntactic Frameworks. In *LACL 2001, Le Croisic, France*, volume 2099 of *Lecture Notes in Computer Science*, p. 17–43.

RAMBOW O., VIJAY-SHANKER K. & WEIR D. (2001). D-tree substitution grammars. *Computational Linguistics*, **27**(1), 87–121.

RETORÉ C. (2000). The Logic of Categorial Grammars. *ESSLI'2000, Birmingham.*

ROGERS J. & VIJAY-SHANKER K. *(1994). Obtaining trees from their descriptions : an application to tree-adjoining grammars.* Computational Intelligence, **10***(4), 401–421.*

SAG I. A., WASOW T. & BENDER E. M. *(2003).* Syntactic Theory : a Formal Introduction. *Center for the Study of Language and INF.*

[7] `www.loria.fr/equipes/calligramme/leopar`

L'abstraction de l'extraction

Jesse TSENG
CNRS, Loria UMR 7503, Vandœuvre-lès-Nancy
tseng@loria.fr

Résumé. L'idée du développement d'un modèle syntaxique de haut niveau pour le français compatible avec plusieurs formalismes syntaxiques semble présupposer que chaque formalisme cible contribue une vision unitaire et stable des analyses à retenir (ou qu'il accepte une telle vision venant du modèle abstrait). Or, il est possible qu'un formalisme admette plusieurs analyses pour un phénomène donné, qui ont toutes le statut d'hypothèses empiriques. L'objectif de cet exposé est de présenter une étude détaillée de l'analyse de l'extraction en HPSG, un domaine caractérisé par une diversité de propositions analytiques et beaucoup de questions ouvertes. Les problèmes évoqués alimenteront une discussion sur le contenu concret du modèle syntaxique abstrait à définir.

Abstract. The idea of developing a high-level syntactic model for French, compatible with several distinct formalisms, seems to presuppose that each target formalism can either contribute a stable set of analyses, or be prepared to incorporate such a set of analyses from the abstract model. It is possible, however, for one formalism to allow several analyses for a given phenomenon, all of which must be considered to be empirical hypotheses. This talk presents a detailed survey of the analysis of extraction in HPSG, an area that has given rise to many proposals and continues to bring up unanswered questions. The points raised here will help advance the discussion of the actual content of the high-level syntactic model that we hope to define.

Mots-clés : extraction, formalismes grammaticaux, grammaires électroniques, HPSG.

Keywords: extraction, grammar formalisms, grammar implementation, HPSG.

1 Introduction

Le développement d'un modèle de description syntaxique de haut niveau, indépendant (dans la mesure du possible) de tout formalisme particulier, implique un travail d'abstraction et de généralisation qui doit s'effectuer d'abord pour chacun des formalismes cibles. Pour certains phénomènes simples, une seule analyse s'impose clairement et recueille le consensus de tous les linguistes et informaticiens travaillant dans un formalisme donné. Mais pour les phénomènes plus complexes un tel consensus s'avère rarissime. Dans la plupart des cas, plusieurs analyses, toutes valables d'un point de vue formel, sont envisageables, et leurs mérites comparés font l'objet de sérieux débats entre spécialistes.

La possibilité d'avoir plusieurs analyses d'un même phénomène dans un formalisme donné est tout à fait souhaitable, puisqu'en tant qu'outil de modélisation linguistique, un formalisme

devrait idéalement permettre la formulation de toutes les hypothèses, afin de les comparer et de les tester.

Cependant, cette multiplicité d'analyses peut rendre plus difficile la comparaison des formalismes, une tâche essentielle qui est, par exemple, au centre des activités du projet Mosaïque. Les objets d'une telle comparaison ne sont pas bien définis, étant donné que suivant le phénomène étudié, chaque formalisme ne présente pas une vision unifiée. Il faut essayer d'identifier ce qui est partagé par toutes les analyses à retenir pour chaque formalisme ; ces efforts constitueront une étape intermédiaire importante dans l'élaboration à terme du modèle abstrait couvrant tous les formalismes.

2 L'extraction

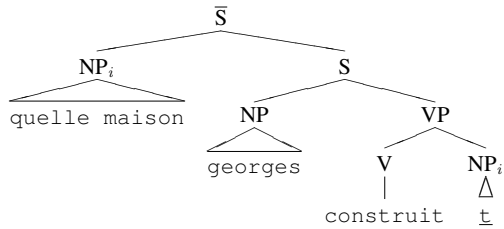
Cette section présente une étude de cas concernant l'analyse de l'extraction dans le formalisme HPSG (la grammaire syntagmatique guidée par les têtes, (Pollard & Sag, 1994)). Nous considérerons l'exemple suivant, correspondant au cas le plus simple (celui d'un verbe transitif dont on extrait le COD) :

(1) Georges construit une maison. $\rightsquigarrow$
   a. Je ne sais pas [quelle maison Georges construit].
   b. Voici la maison [que Georges construit].

Les structures à extraction ont fourni la motivation principale pour l'approche transformationnelle de la syntaxe formelle. Intuitivement, la maison joue le même rôle dans toutes les phrases en (1), et les variantes interrogative et relative peuvent être dérivées de la première phrase en déplaçant les éléments.

Rappelons un peu plus en détails l'analyse transformationnelle de (1a). On construit d'abord une première phrase simple SVO *Georges construit quelle maison* (la structure « profonde »). Dans cette structure, la sous-catégorisation du verbe *construit* est réalisée normalement, avec un COD post-verbal. Mais la phrase ne peut pas se combiner avec le verbe *savoir* sous cette forme : dans l'interrogation partielle indirecte en français, l'antéposition du syntagme interrogatif est obligatoire. Une transformation de déplacement syntaxique doit être appliquée afin de produire la structure de surface suivante :

(2) analyse transformationnelle (arbre « de surface »)



La position du COD post-verbal est toujours occupée dans cette structure, par une « trace » qui a toutes les propriétés d'un NP (elle reçoit notamment le cas accusatif et le rôle sémantique attribués par le verbe *construit*), mais qui n'a aucune réalisation phonologique. Le NP déplacé *quelle*

maison apparaît maintenant en tête de phrase, dans une position non-argumentale (spécifieur de $\overline{\text{S}}$, la projection étendue de S), mais son interprétation syntaxique et sémantique (en tant que COD de *construit*) est récupérable grâce à l'établissement de la « chaîne » transformationnelle reliant les deux positions notées par les indices i.

L'existence d'une opération de déplacement syntaxique semble encore plus évidente quand on considère le caractère non-borné de l'extraction :

(3) Je ne sais pas quelle maison$_i$ [tu dis [que Paul croit [qu'il ne faut pas [que Marie apprenne [que Georges va construire t$_i$]]]]].

L'analyse de tels exemples sans recours à la notion de déplacement a été un défi fondamental pour tous les formalismes non-transformationnels. En HPSG, l'étude de cette question a abouti à un ensemble assez vaste d'analyses différentes. Plusieurs d'entre elles seront présentées schématiquement et commentées dans la section suivante.

Un certain niveau de détail technique sera inévitable dans cette discussion. Nous n'avons pas l'objectif de faire une comparaison minutieuse des toutes les analyses HPSG possibles. Pour les non-spécialistes il suffit de s'assurer que les analyses présentées sont bien distinctes, et valables dans leurs grandes lignes.

3 Déclenchement de la dépendance

Nous nous intéressons d'abord à la structure du groupe verbal `georges construit`, souligné en (1a) et (1b), où se trouve le site de l'extraction (c.-à-d. position canonique du COD extrait).

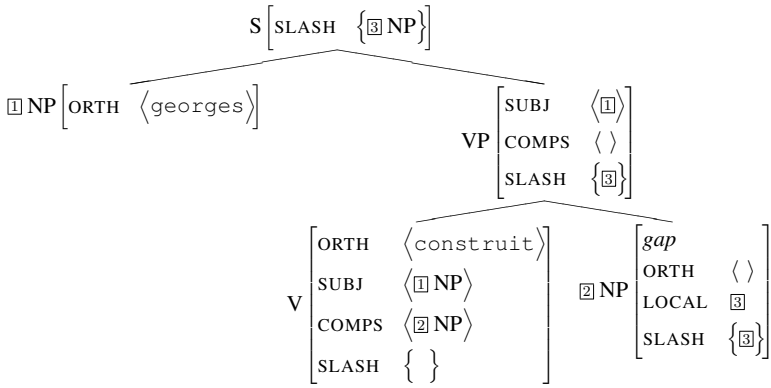
Les deux premières analyses héritent de la notion de « trace » utilisée dans l'analyse transformationnelle pour réaliser explicitement le site de l'extraction dans la structure syntaxique. Comme dans l'arbre (2), la trace est un NP syntaxiquement et sémantiquement, et peut donc apparaître dans toutes les mêmes positions syntaxiques qu'un NP ordinaire, mais elle n'a pas de phonologie/orthographie[1].

L'analyse **Ia** correspond au traitement proposé dans la première partie de (Pollard & Sag, 1994). L'objectif des auteurs à ce stade de l'exposition du modèle était de développer une formalisation non-transformationnelle de l'analyse traditionnelle. Autrement dit, ils voulaient démontrer la possibilité de construire directement une structure de surface correspondant à (2), sans passer par une structure « profonde ».

La première étape de cette analyse est de permettre au verbe `construire` de se combiner directement avec la trace (un item lexical de type *gap*). Aucune règle spéciale n'est nécessaire ; il s'agit d'une structure de complémentation habituelle. Mais la représentation lexicale particulière de la trace (avec un ensemble SLASH non-vide) sert à déclencher la dépendance.

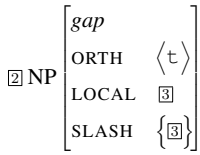
[1]Nous utiliserons ici uniquement l'attribut ORTH(OGRAPHIE), à l'instar de la plupart des grammaires électroniques implémentées

(4) **Analyse Ia** : trace nulle



Il est bien connu que dans une grammaire électronique, l'utilisation de catégories vides ([ORTH ⟨ ⟩]) est problématique. Un analyseur ascendant sans optimisation extra-grammaticale essaiera d'insérer toutes les catégories vides du lexique à toutes les positions syntaxiques, avec des conséquences catastrophiques pour l'efficacité. L'analyse Ib, où la trace est réalisée orthographiquement comme "t", permet d'éliminer ce problème.

(5) **Analyse Ib** : trace réalisée – identique à **Ia**, mais avec

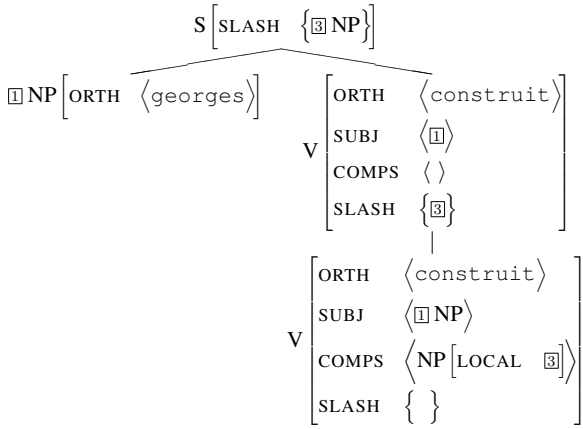


Cette analyse suppose que l'on donne à l'analyseur une entrée « pré-analysée », de la forme "georges construit t" (au lieu de "georges construit", comme en (4)). Évidemment, pour une grammaire destinée au traitement de vrais textes, cette condition est inadmissible. Mais on pourrait choisir d'implémenter l'analyse **Ib**, par exemple, dans une grammaire pédagogique.

Les catégories vides sont problématiques également au plan conceptuel : Les analyses en HPSG sont censées être « surfacistes », et l'utilisation de la trace, qui n'a aucune réalisation en surface, est incompatible avec cette vision. De plus, après les premières affirmations avancées dans la grammaire transformationnelle, la réalité psycholinguistique des traces a été sérieusement remise en question (Sag & Fodor, 1994). L'analyse **Ia** a donc été remplacée très rapidement dans la version standard du modèle HPSG. Certains auteurs continuent, néanmoins, à se servir de traces (notamment (Levine, 2003), pour l'extraction de modifieurs).

Dans les analyses en (4) et (5), le site de l'extraction est représenté dans l'arbre syntaxique. Cette approche a cédé la place aux analyses où l'extraction est déclenchée au niveau lexical, avec une structure syntaxique réduite. De manière générale, les représentations lexicales très riches de HPSG permettent de simplifier ou d'éliminer beaucoup d'opérations traditionnellement réalisées en syntaxe. Dans le dernier chapitre de (Pollard & Sag, 1994), on propose l'analyse **IIa**, dans laquelle la trace n'apparaît plus dans l'arbre.

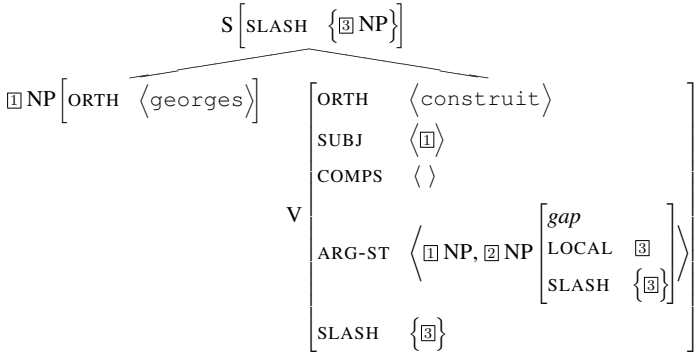
(6) **Analyse IIa** : règle lexicale



Dans l'analyse **Ia**, la présence de la trace avait deux effets : de supprimer le groupe nominal sur la liste COMPS du verbe, et d'introduire un élément correspondant à ce groupe nominal dans l'ensemble SLASH. Dans (6), ces deux opérations s'effectuent directement dans la description lexical du verbe, grâce à une règle lexicale. D'autres règles sont nécessaires pour traiter l'extraction du sujet et des modifieurs ; la suppression de la trace du lexique implique donc un alourdissement des mécanismes lexicaux. L'élimination des traces permet en revanche une simplification des structures syntaxiques. Dans notre exemple, le verbe `construit` ne forme plus de syntagme tête-complément ; il se combine directement, en tant que tête lexicale, avec son sujet `georges`.

Le caractère procédural et non monotone des règles lexicales n'est pas tout à fait compatible avec la vision déclarative de HPSG, bien qu'il soit possible de les intégrer formellement dans le modèle . Pour certains linguistes, le statut théorique des règles lexicales reste problématique ; ils essaient de réanalyser les phénomènes habituellement traités par règle lexicale (par ex. le passif) en utilisant uniquement les outils formels de base de HPSG. (Bouma *et al.*, 2001) propose une alternative à la règle lexicale pour l'extraction dans l'analyse **IIa**.

(7) **Analyse IIb** : sous-spécification de *synsem*

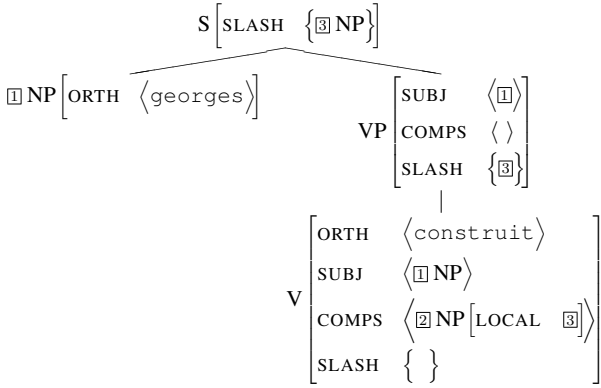


Cette approche exploite le système de typage de HPSG, en supposant une entrée lexicale sous-spécifiée pour le verbe. Cette entrée contient une représentation abstraite de la structure argumentale (ARG-ST) du verbe, valable pour toutes les différentes réalisations possibles des arguments. Le type de l'élément correspondant au COD de `construit` (le deuxième NP sur la liste) peut être résolu de deux façons. S'il est *canonical*, il apparaît également sur la liste COMPS et le COD est réalisé syntaxiquement dans sa position canonique grâce au schéma tête-complément[2]. Si, en revanche, il est de type *gap*, comme dans (7), il n'apparaît pas sur COMPS (qui reste vide, comme pour un verbe intransitif) et il introduit un élément dans l'ensemble SLASH. Le verbe se combine directement avec le sujet `georges`, comme dans l'analyse **IIa**.

Les deux dernières analyses que nous présenterons ci-dessous correspondent à ce que l'on trouve dans beaucoup de grammaires électroniques HPSG, en particulier celles développées avec la « *Grammar Matrix* » (Bender *et al.*, 2002). Elles représentent un retour à l'approche syntaxique, avec une entrée lexicale du verbe qui spécifie une complémentation canonique, mais avec des règles syntaxiques spéciales qui ne réalisent pas cette attente de manière canonique.

Dans la première variante de cette analyse, on définit une règle unaire qui peut s'appliquer au même moment que la règle de complémentation habituelle. Au lieu de réaliser le complément canoniquement, cette règle vide la liste COMPS du verbe et introduit l'élément correspondant dans SLASH.

(8) **Analyse IIIa** : règle de complémentation nulle



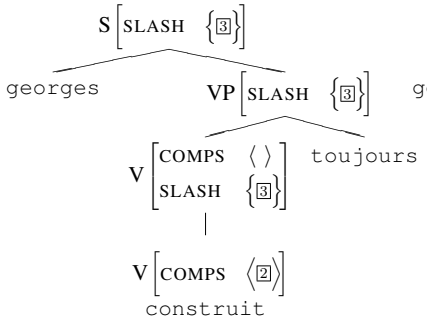
La structure en (8) ressemble beaucoup à celle de l'analyse **IIa** en (6), mais l'ajout d'un adverbe

[2] Voici l'entrée lexicale correspondant à cette résolution de ARG-ST :

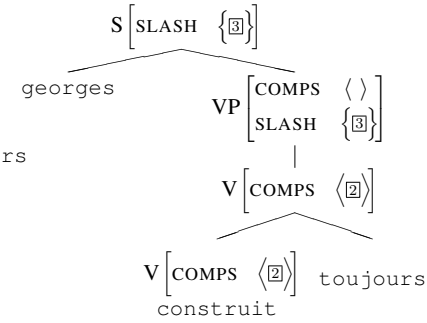
(i) V [ORTH ⟨construit⟩ SUBJ ⟨1⟩ COMPS ⟨2⟩ ARG-ST ⟨1 NP [SLASH 4], 2 NP [*canon* SLASH 5]⟩ SLASH 4 ∪ 5]

permet d'illustrer la différence entre les deux :

(9) a. analyse lexicale IIa

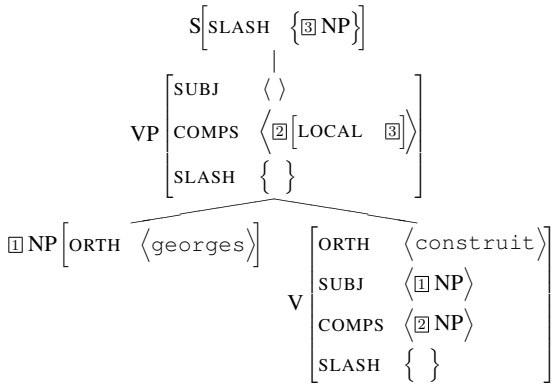


b. analyse syntaxique IIIa



En principe, sans contrainte supplémentaire, dans (9b), la règle syntaxique unaire aurait pu s'appliquer d'abord au verbe `construit`, avant l'adjonction de l'adverbe, donnant un arbre isomorphique à (9a). Et chaque nouvel élément post-verbal dans le VP (complément ou ajout) aggravera cette pseudo-ambiguïté, puisque la règle unaire pourra s'appliquer indifféremment à tous les niveaux, sans incidence sur l'analyse globale. Un moyen de remédier à ce problème serait de reformuler la règle unaire pour qu'elle s'applique uniquement au plus haut niveau de la phrase S, s'il reste un élément de valence non exprimé après la réalisation de tous les autres dépendants syntaxiques :

(10) **Analyse IIIb** : contrainte sur S



Il n'est pas nécessairement utile de retenir tous les détails des six analyses que nous venons d'exposer. Mais chacune (à l'exception de **Ib**, qui suppose la matérialisation artificielle de la trace dans l'orthographe) représente une analyse envisageable de l'extraction en HPSG : formellement valable, linguistiquement justifiable et implémentable dans une grammaire électronique. Même s'il y a des aspects techniques spécifiques au formalisme HPSG, on évoque un certain nombre de questions plus générales :

– la place des catégories vides dans l'analyse syntaxique

– la répartition du travail entre contraintes lexicales et règles syntaxiques
– le statut des règles lexicales
– l'utilité du typage des objets linguistiques.

L'analyse détaillée de la structure interne du syntagme verbal `georges construit` continuera à faire l'objet d'un débat entre les specialistes de HPSG. Mais on peut noter que toutes les analyses arrivent au même résultat global : une projection maximale (valence saturée) du verbe, avec une représentation de l'identité du COD extrait, dans l'ensemble SLASH :

(11)
$$\mathrm{S}\left[\begin{array}{ll} \textsc{head} & \textit{verb} \\ \textsc{val} & \left[\begin{array}{ll} \textsc{subj} & \langle\,\rangle \\ \textsc{comps} & \langle\,\rangle \end{array}\right] \\ \textsc{nonloc} \mid \textsc{slash} & \{\mathrm{NP}\} \end{array}\right]$$

Une stratégie envisageable pour la formulation d'une analyse de haut niveau serait donc de renoncer à définir la structure syntaxique complète des constructions à extraction. L'analyse abstraite préciserait uniquement les propriétés du groupe verbal, sans aller jusqu'au niveau des items lexicaux.

4 Propagation et terminaison de la dépendance

Toutes les analyses étudiées dans la section précédente pour le déclenchement de la dépendance d'extraction sont conçues pour introduire un élément dans SLASH et le faire remonter jusqu'au niveau de la phrase complète. Il faut ensuite rendre compte de la propagation de cet élément SLASH, qui encode l'identité de l'argument extrait. Comme on le sait, la propagation de cette information est en principe non-bornée.

En HPSG, la propagation de SLASH est assurée par le Principe des traits NON-LOCAL. Il existe en gros deux versions de ce principe. La version proposée par (Pollard & Sag, 1994) est syntaxique : dans chaque combinaison syntaxique, l'ensemble SLASH du syntagme dominant est la réunion des ensembles SLASH de toutes les branches (branche tête et branches non-têtes). Dans (Bouma *et al.*, 2001), ce principe est remplacé par une contrainte lexicale (« SLASH Amalgamation Principle – SLAM ») : la tête lexicale rassemble les valeurs SLASH de tous les éléments de sa structure argumentale ARG-ST (voir par ex. la représentation lexicale donnée dans la note 2). Ensuite, dans chaque combinaison syntaxique, c'est la valeur SLASH de la branche tête qui est propagée au niveau supérieur (« SLASH Inheritance Princple – SLIP »). Les autres branches ne peuvent contribuer qu'indirectement, par l'intermédiaire de la tête.

Les différences entre ces deux mécanismes de propagation peuvent être très significatives, surtout dans le traitement des ajouts (qui ne figurent pas, a priori, dans la structure argumentale de la tête). La proposition de (Bouma *et al.*, 2001) semble s'imposer, mais pour des raisons plus théoriques qu'empiriques : les opérations guidées par la tête sont naturellement privilégiées en HPSG. Les conséquences empiriques de ce choix pour l'analyse du français n'ont pas été suffisamment étudiées.

La propagation de SLASH est en fait relativement restreinte en français, par rapport à l'anglais. Les dépendances d'extraction ne sont véritablement non-bornées que dans les enchâssements de compléments phrastiques (3) ou infinitifs :

(12) la maison que [Georges va devoir essayer de réussir à vouloir pouvoir commencer à penser à construire]

En particulier, les chaînes de groupes prépositionnels qui sont possibles en anglais (13) ne le sont pas en français : un seul groupe prépositionnel rend l'extraction agrammaticale (14).

(13) the man that [George is talking to the mother of the brother of the uncle of the best friend of the sister of]

(14) *l'homme dont [Georges connaît la mère du frère] / [Georges parle au frère]

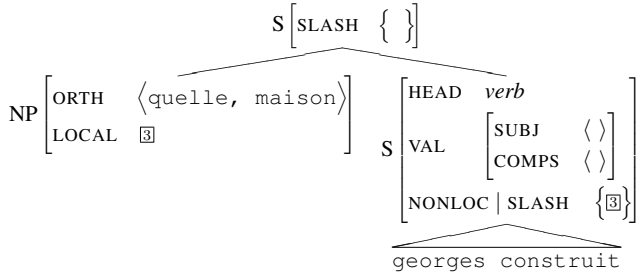
Pour rendre compte de cette restriction en HPSG, il suffit de spécifier que les entrées lexicales des prépositions portent le trait [SLASH { }] (Abeillé *et al.*, 2006)[3]. Mais tous les formalismes grammaticaux ne permettent pas la formulation d'une telle contrainte lexicale. Dans un formalisme de haut niveau on pourrait exprimer la même chose au niveau du groupe prépositionnel : les projections de P en français ne doivent pas contenir un site d'extraction.

Les autres contraintes d'îlot sont à traiter de la même manière. Il faut, par exemple, empêcher la propagation de l'extraction hors d'un sujet, d'une proposition interrogative, d'une circonstancielle, etc. Les travaux linguistiques aboutissent à des formulations assez générales et abstraites de ces contraintes, qui sont ensuite traduites en contraintes HPSG (et parallèlement pour les autres formalismes).

Au sommet de la dépendance, l'identité du syntagme extrait est récupérée dans SLASH, et la propagation de cette information est arrêtée (l'élément correspondant est supprimée de l'ensemble SLASH). L'analyse de cette partie de la construction en HPSG a aussi évolué, et les analyses qui font référence pour l'anglais sont celles présentées dans (Sag, 1997) (pour les relatives) et (Ginzburg & Sag, 2001) (pour les interrogatives). Ces analyses ne sont que partiellement transférables au français, étant donné la syntaxe très particulière des ces constructions. La diversité de structures ne permet pas une approche unifiée.

Nous présentons, à titre d'exemple, une analyse schématique de la proposition interrogative indirecte de la phrase (1a) :

(15) structure tête-antéposé



[3]En supposant l'approche SLAM/SLIP de Bouma *et al*. Sinon, il faut contraindre le complément de la préposition :

(i) *prep-word* → [VAL | COMPS ⟨[SLASH { }]⟩]

Le partage de la valeur LOCAL (notée 3) assure l'identité entre le syntagme interrogatif antéposé et le COD du verbe `construit` extrait. Le syntagme antéposé `quelle maison` reçoit donc la bonne interprétation grammaticale. Au niveau supérieur du syntagme tête-antéposé, l'ensemble SLASH est vide : la propagation de la dépendance s'arrête après la réalisation du syntagme antéposé. La phrase est maintenant complète, et peut servir de complément au verbe `sais` en (1a).

5 Conclusion

L'analyse de l'extraction en HPSG, même pour les cas les plus simples, ne constitue pas un système unifié et stabilisé. On trouve plutôt une multitude d'hypothèses à valider, guidées par un certain nombre de principes, qui eux aussi, peuvent évoluer.

Le formalisme HPSG ne contribue donc pas une vision définie des analyses « correctes » des phénomènes d'extraction. Les résultats des travaux en HPSG dans ce domaine ne fournissent pas directement le contenu du modèle syntaxique de haut niveau. Ils incitent plutôt à réfléchir sur le niveau de détail et d'abstraction que l'on peut souhaiter réaliser en développant ce modèle.

Références

ABEILLÉ A., BONAMI O., GODARD D. & TSENG J. (2006). The syntax of French *à* and *de* : An HPSG analysis. In P. SAINT-DIZIER, Ed., *Syntax and Semantics of Prepositions*, p. 147–162. Dordrecht : Springer.

BENDER E. M., FLICKINGER D. & OEPEN S. (2002). The Grammar Matrix : An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In J. CARROLL, N. OOSTDIJK & R. SUTCLIFFE, Eds., *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, p. 8–14. Taipei.

BOUMA G., MALOUF R. & SAG I. A. (2001). Satisfying constraints on extraction and adjunction. *Natural Language and Linguistic Theory*, **19**, 1–65.

GINZBURG J. & SAG I. A. (2001). *Interrogative Investigations : The Form, Meaning and Use of English Interrogatives*. Stanford, CA : CSLI Publications.

LEVINE R. D. (2003). Adjunct valents : cumulative scoping adverbial constructions and impossible descriptions. In J.-B. KIM & S. WECHSLER, Eds., *Proceedings of the 9th International HPSG Conference*, p. 209–232. Stanford, CA : CSLI Publications.

POLLARD C. & SAG I. A. (1994). *Head-Driven Phrase Structure Grammar*. Stanford, CA : CSLI Publications. Distributed by University of Chicago Press.

SAG I. A. (1997). English relative clause constructions. *Journal of Linguistics*, **33**, 431–484.

SAG I. A. & FODOR J. D. (1994). Extraction without traces. In R. ARANOVICH, W. BYRNE, S. PREUSS & M. SENTURIA, Eds., *Proceedings of the Thirteenth West Coast Conference on Formal Linguistics*, p. 365–384. Stanford, CA : CSLI Publications.

XMG : eXten*ding* MetaGrammars to MCTAG*

Yannick PARMENTIER[1], Laura KALLMEYER[2], Timm LICHTE[2],
Wolfgang MAIER[2]
[1] LORIA–Nancy Université, Campus Scientifique
BP 239, F-54 506 Vandœuvre-Lès-Nancy Cedex, France
[2] SFB 441 - University of Tübingen,
Nauklerstr. 35, D-72074 Tübingen, Germany
parmenti@loria.fr, lk@sfs.uni-tuebingen.de,
{timm.lichte,wo.maier}@uni-tuebingen.de

Résumé. Dans cet article, nous présentons une extension du système XMG (*eXtensible MetaGrammar*) afin de permettre la description de grammaires d'arbres adjoints à composantes multiples. Nous présentons en particulier le formalisme XMG et son implantation et montrons comment celle-ci permet relativement aisément d'étendre le système à différents formalismes grammaticaux cibles, ouvrant ainsi la voie au multi-formalisme.

Abstract. In this paper, we introduce an extension of the XMG system (eXtensible MetaGrammar) in order to allow for the description of Multi-Component Tree Adjoining Grammars. In particular, we introduce the XMG formalism and its implementation, and show how the latter makes it possible to extend the system relatively easily to different target formalisms, thus opening the way towards multi-formalism.

Mots-clés : formalismes syntaxiques, grammaires d'arbres, métagrammaires.

Keywords: syntactic formalisms, tree-based grammars, metagrammars.

1 Introduction

For many NLP applications (*e.g.* generation, machine translation, etc.), large linguistic resources are needed. These resources include (but are not limited to) lexicons and grammars. The latter were originally written by hand. This task of grammar writing was requiring many human resources. Plus, the coherence between the grammatical structures was hard to guarantee (and maintain), as the number of structures / people involved were raising (Erbach & Uszkoreit, 1990). To deal with these issues, several proposals have been made to automatise grammar production. The main proposals are grammar extraction (Xia *et al.*, 2000), grammar inference (Higuera, 2001) and grammar generation. In this paper, we focus on the latter. Grammar generation is based on a formal description of the grammar which is processed to produce a real-size grammar. This technique allows the grammar designer to express linguistic generalisations and to test different representation theories. Grammar generation systems can be divided in two

* This work was carried out during a visit of Yannick Parmentier at the SFB 441, University of Tübingen in January 2007.

main categories, systems based on transformation rules (*e.g. meta-rules* or *lexical rules*) and system based on composition rules (*e.g. metagrammars*).

Transformation rules have been used for many syntactic formalisms such as *Generalized Phrase Structure Grammars* (GPSG), where they are called *meta-rules*. The goal of meta-rules is to build new grammatical structures from existing ones. In unification-based grammars, lexicons are usually defined by associating lexical items with a complex category (represented by a feature-structure). Unlike meta-rules which are applied to grammatical structures, the transformation rules are here applied to lexical entries in order to derive new entries, and are thus called *lexical rules*.

Concerning tree-based grammars, such as *Tree Adjoining Grammars* (TAG), a system of transformation rules has been proposed by (Becker, 1993). In this system, transformation rules are called *meta-rules* and applied to lexical entries containing no longer feature-structures but tree structures (whose nodes may be labelled with feature-structures). In that case, the meta-rules are used to derive new trees. A meta-rule has the following shape : $LHS \rightarrow RHS$, where LHS (respectively RHS) represents the left-hand side (resp. right-hand side) of the rule and consists of a tree fragment. If the left-hand side of the rule matches a given lexical entry, then a tree transformation occurs, replacing the left-hand side in the tree associated with the entry by the right-hand side. For instance, the rule given in Fig. 1 derives the tree fragment for a clitic object starting from the tree fragment for the canonical nominal object.

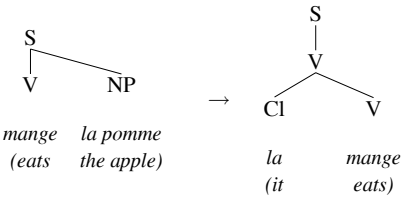


FIG. 1 – Transformation rule for Clitic-Object in French with TAG.

One drawback of such a system comes from the fact that the rule applications must be controlled to avoid over-generation and infinite loops. (Prolo, 2002) proposes to use a declaration (for each lexical entry) of valid application orderings as a control process. Nevertheless, when dealing with real-size grammars, this task of rule ordering may be tedious.

The second trend in grammar generation, *i.e.* systems based on composition rules, has emerged from works on tree-based grammars, especially TAG (Candito, 1996). Here, the factorisation needed to describe a real-size grammar is not provided by transformation rules allowing for the *expansion* of canonical trees. Instead, the factorisation arises from the definition of (i) elementary tree fragments, and (ii) composition rules over these fragments. This factorised definition of the grammatical structures is sometimes called a *metagrammar*. Several metagrammatical systems have been developed, especially for TAG[1]. Among these, one may cite *eXtensible MetaGrammar* (XMG), which distinguishes itself from previous approaches by its extensibility and flexible management of variable scopes (Duchier *et al.*, 2004). On top of providing a high-level language allowing for the description of TAG grammars, the XMG language can be extended to deal with other grammatical formalisms. Such an extension would make it easier to study the

[1] See (Duchier *et al.*, 2004) for a comparison of these systems.

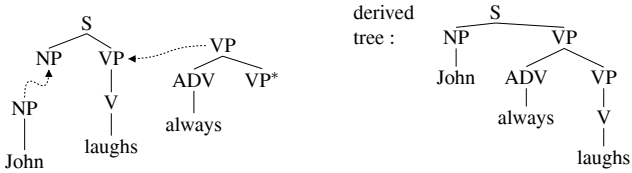


FIG. 2 – TAG derivation for *John always laughs*

common points between *meta*-descriptions for different formalisms, and opens the way towards *multi-formalism*, which is one of the targets of the Mosaïque project[2]. To illustrate this extensibility, we take the example of *Multi-Component Tree Adjoining Grammars* (MCTAG). The paper is organised as follows. In section 2, we introduce MCTAG, motivate their use considering the description of German, and point out the limitations of TAG metagrammatical systems with respect to the description of MCTAG. Then in section 3, we present the XMG formalism and its implementation. In section 4, we show how, concretely, the XMG system has been extended to support the description of MCTAG. Finally, in section 5, we conclude and point out some perspectives for future work.

2 Multi-Component Tree Adjoining Grammars (MCTAG) and "Meta" MCTAG

Tree Adjoining Grammars (TAG) as originally defined by (Joshi *et al.*, 1975) consist of elementary trees which can be combined via *substitution* (replacing a leaf with a new tree) and *adjunction* (replacing an internal node with a new tree). In case of an adjunction, the tree being adjoined has exactly one leaf that is marked as the foot node (marked with an asterisk). Such a tree is called an *auxiliary* tree. When adjoining it to a node n, in the resulting tree, the sub-tree with root n from the old tree is attached to the foot node of the auxiliary tree. Non-auxiliary elementary trees are called *initial* trees. A derivation starts with an initial tree. In a final derived tree, all leaves must have terminal labels. For a sample derivation see Fig. 2.

An extension of TAG that has been shown to be useful for several linguistic applications is Multi-Component TAG (MCTAG) (Joshi, 1987; Weir, 1988). An MCTAG additionally lets one declare tree sets consisting of elementary trees, meaning two things : firstly, using a tree set implicates using all the trees belonging to it ; secondly, the attachment (i.e. adjunction or substitution) of the trees of a tree set can be restricted with respect to the place of attachment : if the trees of a tree set are attached to the same elementary tree, the MCTAG is called *tree-local* ; if they are attached to the same tree set, the MCTAG is called *set-local* ; otherwise (i.e. without attachment restriction) the MCTAG is called *non-local*. Tree-local and set-local MCTAG are polynomially parsable (the former are even strongly equivalent to simple TAG) while non-local MCTAG are NP-complete (Rambow & Satta, 1992).

From a linguistic point of view, MCTAG are particularly interesting when modelling long distance dependencies and movement effects, since the notion of tree sets allows for a conjoint

[2] *Cf.* `http://mosaique.labri.fr`

representation of fillers and gaps in one lexical entry. An early application of tree-local MCTAG was thus dedicated to the modelling of extraposed relative clauses (Kroch & Joshi, 1987). Another field of application is the modelling of scrambling data in German, which furthermore necessitate a TAG formalism more expressive than simple TAG or tree-local MCTAG. Therefore (Rambow, 1994) has provided an MCTAG variant, called V-TAG, that is in fact non-local. However, the desired computational properties are re-implemented via the use of dominance constraints and the admission of non-synchronous attachments of the trees of a tree set. More recently, (Kallmeyer, 2005) has developed another MCTAG variant in order to account for the scrambling data, namely MCTAG with shared nodes (SN-MCTAG). Other than V-TAG, the notion of SN-MCTAG is built upon tree-local MCTAG, but the notion of locality is relaxed, such that non-local attachments are permitted under certain circumstances. Kallmeyer shows that SN-MCTAG is still tractable in polynomial time. As an example a tree-local MCTAG analysis of scrambled constituents in the German Mittelfeld is provided in Fig. 3. Note, that the analysis is the same when using SN-MCTAG.

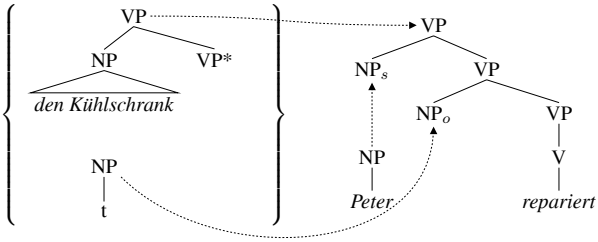


FIG. 3 – A tree-local MCTAG analysis of the German sentence "*[dass] den Kühlschrank Peter repariert*" ('[that] Peter repairs the fridge'). The object noun *den Kühlschrank* ('the fridge') is fronted, leaving a trace in its base position behind the subject NP.

The different MCTAG variants proposed in the literature can be distinguished with respect to the derivations they licence. But independent from this, they all consist of sets of elementary TAG trees. The original metagrammar system XMG did not support tree sets, which made it difficult to use for encoding MCTAG. Indeed, when designing metagrammars for TAG, the user defines tree descriptions whose models are TAG trees. The only grouping of these trees is made through the way the tree descriptions are gathered[3]. In practice, the descriptions are gathered with respect to *sub-categorisation frames* (Crabbé, 2005). In the case of MCTAG, one may want to describe sets of trees according to specific criteria (defined by the metagrammar designer). The XMG language has thus been extended so that the metagrammar designer can define tree descriptions whose models are sets of trees.

3 XMG : an extensible metagrammatical framework

In this section, we introduce the XMG system, and present its main features making it extensible. This introduction will be followed (next section) by a step-by-step presentation of how to extend it to different grammatical formalisms, taking the example of MCTAG.

[3]This gathering is specified using conjunctions and disjunctions, *cf* section 3.

By XMG, we refer to both (i) a metagrammatical formalism, *i.e.* a formal language allowing to express abstractions over the structures of a grammar, and (ii) an implementation of this formalism, in other words, a compiler for the XMG language.

3.1 The XMG formalism.

Definition of elementary tree fragments. The XMG language allows to describe reusable tree fragments through abstractions called *classes*. A class corresponds to the association of a name with a content :

$$\textit{Class} \quad ::= \quad \textit{Name} \rightarrow \textit{Content} \tag{1}$$

For tree-based grammars, this content corresponds to a tree fragment (*Content*::=*Description*) represented using a tree description logic formula built on the following language :

$$\begin{aligned} \textit{Description} \ ::= \ & x \rightarrow y \mid x \rightarrow^{+} y \mid x \rightarrow^{*} y \mid x \prec y \mid x \prec^{+} y \mid x \prec^{*} y \mid \\ & x[f{:}E] \mid x(p{:}E) \mid \textit{Description} \wedge \textit{Description} \end{aligned} \tag{2}$$

where x, y are node variables, $\rightarrow$ represents the dominance relation (*mother-of* relation), $\rightarrow^{+}$ its transitive closure, and $\rightarrow^{*}$ its reflexive and transitive closure. $\prec$ refers to the precedence between nodes (*sister-of* relation), $\prec^{+}$ its transitive closure, and $\prec^{*}$ its reflexive and transitive closure. $x[f{:}E]$ is the association of the feature f and the value E to the node referred to by the x variable. $x(p{:}E)$ is the association of the property p and the value E to the x node. Note E is an expression and can correspond to either a variable, a constant or a disjunction over constants (so-called *atomic disjunction*).

Definition of combination of tree fragments. Once elementary tree fragments have been defined, it is possible to define combinations over these, using two operators, namely *conjunction* and *disjunction*. Concretely, the XMG language is extended with the following definition :

$$\textit{Content} \quad ::= \quad \textit{Description} \mid \textit{Name} \mid \textit{Content} \vee \textit{Content} \mid \textit{Content} \wedge \textit{Content} \tag{3}$$

The content of a class can either be a description (tree description logic formula) or a name (class instantiation), or a disjunction / conjunction of contents.

This part of the XMG language allows for a flexible control over the class combinations, making it possible to express linguistic properties of natural languages. For instance, the fact that transitive verbs are made of a subject, a verbal morphology (active or passive) and a object can be described within XMG as illustrated below :

$$\textit{transitive} \quad \rightarrow \quad \textit{subject} \ \wedge \ \textit{morphology} \ \wedge \ \textit{object}$$

To illustrate the expressive power allowed by XMG[4], you can imagine that a subject is not a single tree fragment but a disjunction of tree fragments, each one defining a syntactic realisation of the subject.

An important remark has to be made here. In the above example, nothing is said about the way the tree fragments are "stuck" together, *i.e.* about how nodes are identified. This node identification was a central point in previous metagrammar approaches. In the XMG approach, the scope of a node variable is *by default* local to the class (*i.e.* to the fragment). This means that

[4]Another illustration of this power is the case of the agentless passive. Unlike (Candito, 1996), it does not need any description removal with this language, the description is thus fully declarative.

you can reuse the same node variable in different fragments without any name conflict. When you want to declare that two node variables introduced in different fragments denote the same node, you can use a prefix notation and a node equation :

$$S = subject \ \wedge \ A = morphology \ \wedge \ S.X = A.X$$

Extension to different levels of description. Up to now, we have seen how to factorise syntactic information (tree structures) within a metagrammatical description in the XMG language. In order to allow for the extension of different levels of description (such as semantics, or non-tree-based syntax, *etc.*), we have to extend the XMG language. This extension corresponds to the concept of *dimensions*. The content of a class is a description belonging to a given dimension, each dimension has its own sub-language. Definition (3) is extended by :

$$\begin{array}{lll} \textit{Content} & ::= & \textit{Dimension} += \textit{Description} \mid \textit{Name} \mid \\ & & \textit{Content} \vee \textit{Content} \mid \textit{Content} \wedge \textit{Content} \end{array} \quad (4)$$

For instance, XMG integrates a semantic dimension allowing for the description of predicative formulae.

3.2 The XMG compiler

The language introduced above is processed by a compiler in order to produce the grammar described. Before presenting the architecture of this compiler, we can recall that the XMG language is made of two devices :

– a collection of description languages (one for each dimension), allowing for the description of basic units,
– a combination language.

This duality of the language is reflected within the architecture of the compiler, which performs two main tasks : (a) accumulating basic units (in other words, processing the combination rules), and (b) applying a processing on the accumulated units (for instance, once partial tree descriptions are accumulated, computing the corresponding tree models). While the first task is common to all dimensions, the second task is dimension-dependent.

Processing of the combination rules. It is worth noticing that the XMG combination language corresponds to a *Definite Clause Grammar* (DCG) (Pereira & Warren, 1980). Indeed, when considering tree descriptions as words, conjunctive and disjunctive rules are just DCG rules. This is why the combination language is processed the same way as a DCG would be by a PROLOG compiler. In the DCG paradigm, one has to define axioms, from which PROLOG computes the corresponding DCG parses. In our case, the metagrammar designer also defines axioms (*i.e.* the classes that encode combination rules leading to total tree descriptions, such as $transitive$ in the above example) using the $value$ keyword. In order to have full control on unification, we decided to develop our own WAM-based virtual machine (see (Duchier *et al.*, 2004)). This virtual machine distinguishes between the different dimensions, thus as an output, it produces a list of accumulated descriptions (total tree descriptions for syntax, list of predicates for semantics).

Additional processing of the accumulated descriptions. After the processing of the combination rules, we have a list of descriptions (one description per dimension for each axiom). Let us call this list $L(x) = (D_1, \ldots, D_n)$, where x is an axiom (class name) and D_i the description of the dimension i. Each dimension i is processed by a specific solver S_i, whose role is to produce the models $S_i(D_i)$ of the description D_i.

For the syntactic dimension, say dimension 1, D_1 is a tree description, and the solver S_1 computes all minimal tree models satisfying D_1. Let us briefly introduce S_1. S_1 has been developed as a *Constraint Satisfaction Problem*. The idea behind this is to associate each node variable x of the description D_1 with an integer j, then to define the position of this node in a model as a 5-tuple $N_j^x = (Eq, Up, Down, Left, Right)$ where Eq refers to the node variables (integers) that are identified with x in a model, Up to the node variables that denote the ancestors of x in a model, $Down$ its descendants, etc. Finally the relations between node variables in D_1 are translated into constraints over these 5-tuples, *i.e.* constraints over sets of integers (see (Duchier *et al.*, 2004; Le Roux *et al.*, 2006) for more details).

For the semantic dimension, say 2, D_2 is a list of predicates. There is no need for further processing of this dimension, so S_2 is just the identity operation.

4 Towards a library of operational constraints for describing different target formalisms

Before the work described here took place, the XMG system was supporting the description of TAG, Interaction Grammars (IG)[5], and Hole Semantics. We extended it so that one may also describe MCTAG. This extension was made possible by the modular architecture of the system as advocated in (Le Roux *et al.*, 2006). This extension was made in two steps :

1. extension of the XMG language (either by defining a new dimension with its own sublanguage, or by extending an existing one),
2. definition of the solver for this new / extended dimension.

Extension of the XMG language. As presented above, MCTAG is an extension of TAG in which the elementary structures of the grammar are sets of trees. In a metagrammatical context, the factorisation of an MCTAG corresponds to the definition of tree fragments that are combined to produce (no more trees but) sets of trees. Concretely, this means that we can keep the same tree description language as the one for TAG given in definition (2). Thus, we do not need a new dimension, we can extend the existing syntactic dimension by adding a unary operator to distinguish between descriptions whose models are trees and those whose models are sets of trees. As introduced in the preceding section, the metagrammar designer defines *axioms* (class names) indicating the classes which refer to total descriptions. These axioms are the starting point for the processing of the combination rules. In our extension, we define a second type of axioms using the $setvalue$ keyword. The classes referred to by these new axioms have to be interpreted as descriptions of sets of trees. This means that we have to define a new solver for these descriptions of sets. Thus, we defined a S_3 solver taking as an input a description belonging to dimension 1 (syntax). While $S_1(D_1)$ computes trees, $S_3(D_1)$ computes sets of trees. Note that S_1 and S_3 can be used within the same metagrammar (*i.e.*, share the same tree fragments).

Definition of a solver for MCTAG. While the S_1 solver introduced above applies tree-specific constraints on models (such as the uniqueness of the root node), the S_3 solver we defined for MCTAG behaves differently. S_3 has two major differences compared with S_1[6]. First, there is no

[5]Both TAG and IG were using the same syntactic dimension, as these formalisms are both based on trees.

[6]For lack of space, we do not present neither S_3 nor S_1 in detail here (see (Duchier *et al.*, 2004; Le Roux *et al.*, 2006) for a detailed introduction to S_1).

constraint of root uniqueness, that is to say, two nodes of a model can be such that there is no node above them :

$$\exists\, j,k \in [1..n] \mid N_j^x.Up = \emptyset \wedge N_k^y.Up = \emptyset \wedge (N_j^x.Eq \cap N_k^y.Eq) = \emptyset$$

(n represents the number of node variables in the description). Secondly, two different nodes of a model can belong to two different trees. For our 5-tuple representation, this means that possibly none of their position features (Eq, Up, $Down$, $Left$, $Right$) intersects :

$$\exists\, j,k \in [1..n] \mid (N_j^x.Eq \cup N_j^x.Up \cup N_j^x.Down \cup N_j^x.Left \cup N_j^x.Right) \cap (N_k^y.Eq \cup N_k^y.Up \cup N_k^y.Down \cup N_k^y.Left \cup N_k^y.Right) = \emptyset$$

To illustrate the difference between S_1 and S_3, consider the description A of Fig. 4. This description can be interpreted either as trees ($S_1(A)$), or as sets of trees ($S_3(A)$).

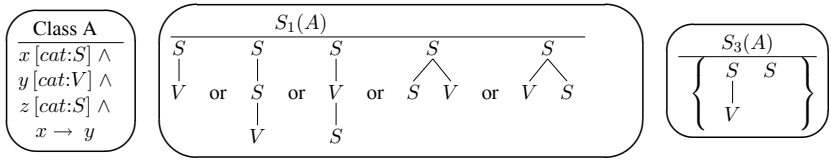


FIG. 4 – Description solving as trees / sets of trees.

When computing trees, S_1 searches for all *minimal* tree models to the description. That is, S_1 does not add any node on top of those referred to by a node variable in the description. S_1 successively tries to identify nodes (when the feature structures labelling the nodes unify), or to add a dominance relation between a node and a local root. S_3 searches for models of sets of trees, more precisely, S_3 does not add any dominance relation on local roots. We have integrated the S_3 solver in the current XMG compiler[7], and we started implementing a metagrammar for German using this new multi-component dimension.

Note that when the metagrammar designer specifies a precedence (or dominance) relation between nodes belonging to distinct trees of a set, S_3 is unable to compute a solution. Such relations between nodes of elements of a set are used in some extensions of MCTAG, and corresponds to (unsolved) constraints on the trees of the set. These constraints have to be applied during parsing. It would be interesting to extend the XMG language to include node relations that are not to be solved.

Towards multi-formalism. We have seen a first extension of the XMG system to deal with MCTAG. As this formalism is based on trees, we did not need a new dimension. Nevertheless, it is worth noticing that such an extension was facilitated by the modular architecture of the system (virtual machine processing DCG rules and solver computing grammatical structures). To sum up, it is possible to extend XMG for compiling a specific dimension provided you define a language for describing it, and a solver for interpreting the corresponding description. The latter can be seen as a set of operational constraints applied to a description in order to produce valid structures (with respect to grammatical criteria).

[7] In the XMG-Tuebingen development branch of the subversion repository, see `http://sourcesup.cru.fr/xmg`.

The next step in this work is to define a library of solvers applying specific operational constraints on grammatical descriptions. These constraints will be selected dynamically by the metagrammar designer depending on the targeted grammatical formalism. Such a library would allow the metagrammar designer to abstract away from the technical aspects of a given grammatical formalism, providing him with a high-level description language. Furthermore, the metagrammar designer would be able to define a linguistic description that would be interpreted by different solvers to produce grammars for different formalisms (*i.e.*, multi-formalism).

A second interesting perspective consists of the development of a device allowing for solver specification. Thus the linguist would be able to define its own grammatical criteria from which the corresponding solver would be generated automatically.

Finally, another interesting perspective concerns parsing directly from the metagrammatical descriptions (*i.e.*, without computing the elementary units of the grammar). This path is followed by the MGCOMP system (Villemonte de la Clergerie, 2005).

5 Conclusion and Perspectives

This paper addresses a central problem of large coverage grammar implementation, namely the difficulty to keep the grammar consistent across its different parts in spite of the considerable redundancy that arises with the increasing size of an electronic grammar. This problem is particularly prominent in lexicalised grammars that do not allow to formulate linguistic generalisations outside the lexical entries. As a solution, eXtensible MetaGrammar (XMG) provides a platform for grammar development that allows to factorise the lexical entries of a grammar into smaller pieces that can then be used in different places. XMG is intended for lexicalised tree grammars, in particular Tree Adjoining Grammars (TAG). TAG allows to describe a large range of linguistic phenomena, in some cases however its expressive power is too limited. One such example is the phenomenon of scrambling in so-called free word order languages such as German. The different proposals for extending TAG in order to account for German scrambling data all have in common that they use an MCTAG, i.e., a grammar consisting of sets of trees instead of trees. The goal of this paper was to extend XMG so that it can be used not only to describe TAG but also MCTAG.

In the paper, we have achieved such an extension by (optionally) relaxing the conditions on the models, in particular omitting the assumption about the uniqueness of the root node. In that case, a minimal model for a given description in the metagrammar might still be a tree (if all nodes are connected in the description) but it could also be a set of disconnected trees. We think that our technique of relaxing the restriction of XMG to tree models opens up interesting perspectives for future work oriented towards other grammar formalisms. The idea to allow graphs different from proper trees as models could be exploited for example for formalisms involving feature structures instead of trees.

Such an extension of metagrammars to different target formalisms would allow to study how the factorisation and the expression of linguistic generalisation is represented in these formalisms (this comparative task would be facilitated by using the same language and system). The results of such a study would make it possible to go further towards *strong* multi-formalism, that is to say towards the compilation of different grammars (*i.e.* in different formalisms) starting from a single meta-description.

Références

BECKER T. (1993). *HyTAG : A new Type of Tree Adjoining Grammars for Hybrid Syntactic Representation of Free Word Order Language*. PhD thesis, Universität des Saarlandes.

CANDITO M. (1996). A principle-based hierarchical representation of LTAGs. In *Proceedings of COLING'96, Kopenhagen*.

CRABBÉ B. (2005). *Représentation informatique de grammaires fortement lexicalisées : Application à la grammaire d'arbres adjoints*. PhD thesis, Université Nancy 2.

DUCHIER D., LE ROUX J. & PARMENTIER Y. (2004). The Metagrammar Compiler : An NLP Application with a Multi-paradigm Architecture. In *2nd Internationale Conference of Mozart/Oz users (MOZ'2004)*, Charleroi.

ERBACH G. & USZKOREIT H. (1990). *Grammar Engineering : Problems and Prospects – Report on the Saarbrücken Grammar Engineering Workshop*. Rapport interne 1, Saarbrücken, Germany.

HIGUERA C. D. L. (2001). Current trends in grammatical inference. *Lecture Notes in Computer Science*, **1876**.

JOSHI A. K. (1987). An introduction to tree adjoining grammars. In A. MANASTER-RAMER, Ed., *Mathematics of Language*, p. 87–114. John Benjamins, Amsterdam.

JOSHI A. K., LEVY L. S. & TAKAHASHI M. (1975). Tree adjunct grammars. *Journal of Computer and System Science*, **10**, 136–163.

KALLMEYER L. (2005). Tree-local multicomponent tree adjoining grammars with shared nodes. *Computational Linguistics*, **31 :2**, 187–225.

KROCH A. S. & JOSHI A. K. (1987). Analyzing extraposition in a tree adjoining grammar. In G. J. HUCK & A. E. OJEDA, Eds., *Discontinous Constituency*, number 20 in Syntax and Semantics, p. 107–149. Academic Press, Inc.

LE ROUX J., CRABBÉ B. & PARMENTIER Y. (2006). A constraint driven metagrammar. In *Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG+8), Sydney, Australia*.

PEREIRA F. & WARREN D. (1980). Definite clause grammars for language analysis —a survey of the formalism and a comparison to augmented transition networks. *Artificial Intelligence*, **13**, 231–278.

PROLO C. A. (2002). Generating the XTAG English grammar using metarules. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'2002)*, p. 814–820, Taipei, Taiwan.

RAMBOW O. (1994). *Formal and Computational Aspects of Natural Language Syntax*. PhD thesis, University of Pennsylvania, Philadelphia. IRCS Report 94-08.

RAMBOW O. & SATTA G. (1992). Formal properties of non-locality. In *Proceedings of 1st International Workshop on Tree Adjoining Grammars*.

VILLEMONTE DE LA CLERGERIE E. (2005). DyALog : a tabular logic programming based environment for NLP. In *Proceedings of CSLP'05*, Barcelona.

WEIR D. J. (1988). *Characterizing Mildly Context-Sensitive Grammar Formalisms*. PhD thesis, University of Pennsylvania.

XIA F., PALMER M. & JOSHI A. (2000). A Uniform Method for Grammar Extraction and Its Application. In *Proceedings of 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Les constructions à verbe support en TAG : intégration à la métagrammaire des verbes pleins du français

Sébastien BARRIER
Laboratoire LLF - Université Paris 7 Denis Diderot,
2 Place Jussieu, 75251 PARIS CEDEX 05
sbarrier@linguist.jussieu.fr

Résumé. Le travail que nous présentons dans cet article montre comment deux hiérarchies différentes peuvent être fusionnées pour décrire la grammaire d'une langue. Jusqu'à présent, la grammaire FTAG était constituée de 2 hiérarchies indépendantes : une pour les verbes pleins et une autre pour les constructions à verbe support. Malgré le fait que leurs descriptions étaient partiellement similaires, aucun effort n'avait été fait pour fournir une description unique. Après avoir effectué un rapide aperçu du traitement des constructions à verbe support dans la grammaire TAG du français, cet article résume les modifications et extensions à apporter aux classes de la hiérarchie des verbes pleins pour prendre en compte les constructions à verbe support.

Abstract. The work we describe in this paper is intended to present how two metagrammatical hierarchies can be merged into one. So far, the FTAG grammar was made of two independent hierarchies : one for main verbs and another one for support verbs. Although their descriptions were pretty similar, no work had been made to design a unique description. This paper shows how the main verb hierarchy can be extended to take into account the description of support verbs. This entails slight changes in the way the MG formalism will be used.

Mots-clés : grammaires d'arbres adjoints, français, noms prédicatifs, verbes supports, métagrammaire, implémentation.

Keywords: tree-adjoining grammars, french, predicative noun, support verb constructions, metagrammar, implementation.

1 Introduction

Depuis maintenant une dizaine d'années, le concept de Métagrammaire (dorénavant MG) est régulièrement utilisé en linguistique pour modéliser des grammaires TAG de taille conséquente (voir par exemple pour le français (Candito, 1999) (Crabbé, 2005) et (Barrier, 2006)). En effet, les MG se révèlent un formalisme de description syntaxique (à vocation multilingue) relativement puissant et permettent ainsi une représentation élégante et en principe non redondante de l'information, dans la mesure où elles utilisent un partage par héritage et s'appuient sur des domaines linguistiques bien définis représentant notamment la sous-catégorisation (dimension 1), les changements de diathèse (dimension 2), et les réalisations de surface des fonctions finales

(dimension 3)[1] . De fait, les MG constituent un niveau de description syntaxique plus abstrait qu'une simple grammaire, et leur utilisation assure au descripteur la formation d'arbres élémentaires corrects, de par la combinaison d'unités arborescentes d'ordre supérieur.

Malheureusement, l'utilisation qui en a été faite jusqu'à présent s'est principalement limitée à des problèmes spécifiques, de sorte qu'aucune généralisation n'a jamais vraiment été proposée[2]. On s'est ainsi bien plus souvent intéressé au réseau hiérarchique qu'au contenu des classes même, et en pratique, un grand nombre d'informations redondantes ont été encodées au sein des classes de dimension 2 et 3, ce qui a empêché un développement global. Le but de cet article est de présenter les modifications à apporter à la métagrammaire des verbes pleins pour pouvoir y intégrer la métagrammaire des verbes supports. Jusqu'à présent, deux hiérarchies distinctes avaient été développées en se fondant sur le langage de description mis au point par (Candito, 1999). Les arbres générés séparément avaient été intégrés au sein d'une unique grammaire, pour former la grammaire FTAG actuelle. Pourtant, il semblait nécessaire de disposer d'une description commune : d'une part, parce que cela permettait à la grammaire d'accroître encore sa cohérence, et d'autre part, parce que ce couplage facilitait le développement et la maintenance de la grammaire.

L'article se compose de 2 parties : la première, plus descriptive, illustre les spécificités de la grammaire TAG des verbes supports par rapport à celle des verbes pleins, la seconde, plus technique, esquisse les modifications à apporter pour former une hiérarchie unique prenant en compte les constructions à verbe plein et les constructions à verbe support.

2 Différentiel des représentations en TAG

Nous rappelons ici brièvement les choix théoriques et linguistiques qui sous-tendent la grammaire TAG du français au niveau de la description des familles à verbe support et à nom prédicatif. Afin de bien en cerner les apports, un bref différentiel est effectué avec les familles des verbes pleins. Bien entendu, il n'est pas question pour nous de dresser un portrait complet de cette grammaire : nous mettons simplement en avant dans cette section les caractéristiques saillantes qui nous paraissent essentielles dans la description d'une métagrammaire unifiée et renvoyons le lecteur à (Abeillé, 1991; Abeillé, 2002) (Candito, 1999) et (Barrier, 2006) pour davantage de détails[3]. Le lecteur non familier avec le formalisme TAG pourra également se reporter à (Abeillé & Rambow, 2000).

[1] Une quatrième dimension permettant de rendre compte de l'ordre des arguments entre eux a été ajoutée par (Barrier, 2006). Cette dimension n'est ici pas pertinente pour notre propos, car elle est identique dans la MG des verbes pleins et des verbes supports.

[2] On peut citer à cela une exception : pour représenter la sous-catégorisation des adjectifs, (Crabbé, 2005) réutilise les descriptions utilisées pour la représentation des verbes pleins, et introduit une nouvelle description pour représenter, « non plus la forme du verbe, mais plutôt la forme de l'adjectif ». Cette réutilisation est facilitée par les choix linguistiques opérés, l'adjectif formant avec le verbe un noyau verbal. L'implémentation est réalisée avec XMG, mais pourrait également être réalisée avec le compilateur de Candito.
Les différences entre XMG et le compilateur de Candito reposent principalement sur 3 points :
- Organisation de l'information : elle est décrite de manière monotone dans XMG ;
- Gestion des variables : elles ne sont pas globales dans XMG ;
- Combinaison des unités arborescentes : XMG utilise un langage de contrôle, alors que le compilateur de Candito repose sur l'utilisation d'un algorithme de croisement qui fixe les domaines d'information des descriptions.

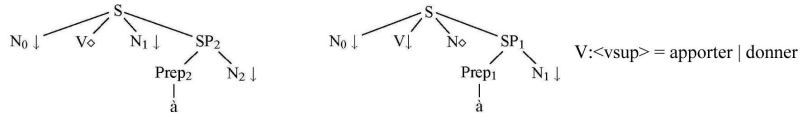
[3] Le lecteur pourra également consulter (Crabbé, 2005) pour une description alternative d'une grammaire TAG du français.

2.1 Aperçu général

S'inscrivant dans le cadre de la grammaire FTAG, les familles des verbes pleins et des verbes supports s'appuient sur des choix linguistiques et théoriques relativement semblables.

Les verbes pleins sont considérés comme des prédicats à projection phrastique sélectionnant tous leurs arguments au même niveau syntagmatique. Dans leur construction canonique, ils sont représentés par des arbres de racine S avec un nœud feuille pour chacun de leurs arguments.

Les verbes supports, dépourvus de sens prédicatif, sont quant à eux, substitués dans des arbres phrastiques à ancre nominale. La sélection du verbe support est réalisée par un trait *vsup* qui indique les formes autorisées par le nom prédicatif[4].

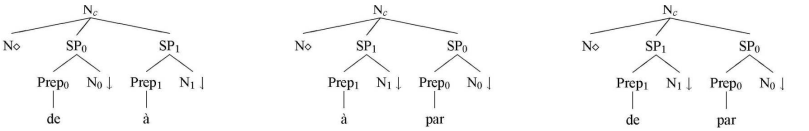


Représentation d'un verbe plein bi-transitif
Max donne un livre aux enfants

Représentation d'un verbe support
Le gouvernement donne son adhésion à ce projet

FIG. 1 – Représentation des verbes pleins et des verbes supports dans la grammaire FTAG

Outre ces constructions verbales, les noms prédicatifs ancrent également des syntagmes nominaux complexes. La relation systématique et paraphrastique entre la construction à groupe nominal et la construction à verbe support correspondante est ainsi assurée. Ces syntagmes nominaux peuvent être à interprétation active, ou passive comme l'illustre la figure suivante[5] :



Une adhésion claire du gouvernement à ce projet (serait idéale)

Une adhésion claire à ce projet par le gouvernement (serait idéale)

Une adhésion claire de ce projet par le gouvernement (serait idéale)

FIG. 2 – Quelques groupes nominaux complexes

2.2 Quelques représentations spécifiques

Notre représentation tient compte également des différentes diathèses déjà utilisées pour les verbes pleins à savoir celles des constructions passives, moyennes, réfléchies, impersonnelles

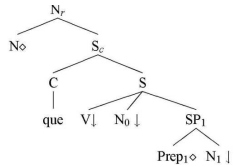
[4]Cette description différencie donc clairement tête syntaxique (verbe support) et tête lexicale (nom prédicatif). Les états antérieurs de la grammaire FTAG considéraient le verbe support et le nom prédicatif comme des co-têtes. La distinction avec les expressions figées n'était pas immédiatement évidente.
La grammaire TAG de l'anglais considère également que le verbe support et le nom prédicatif forment des co-têtes. Mais à la différence du français, les familles des verbes supports se distinguent davantage des familles à verbes pleins. En particulier, les possibilités d'extraction sont beaucoup plus limitées et la passivation n'est pas régulière.

[5]Les exemples de la figure 2 sont adaptés de (G. Gross 1989). L'acceptabilité de la phrase *Une adhésion claire à ce projet par le gouvernement serait idéale* peut varier selon les locuteurs, mais elle est facilitée par l'utilisation du déterminant indéfini.

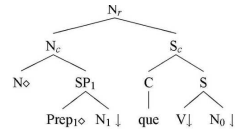
et causatives. Elle présente en revanche certaines particularités : les arbres des familles à verbe support soulignent en effet le comportement double de leurs compléments prépositionnels et phrastiques (c'est-à-dire leur faculté à pouvoir être considérés soit comme complément du verbe soit comme complément du nom (M. Gross 1976)), et permettent le traitement du passif lexical (avec les supports converses (G. Gross 1989)) et du causatif synthétique (avec les opérateurs causatifs (M. Gross 1981)).

Dans le cas général, la grammaire TAG des verbes supports n'adopte pas de double analyse au sens strict : nous avons choisi de considérer les compléments prépositionnels et phrastiques soit comme des compléments du verbe, soit comme des compléments du nom, sans qu'il y ait pour autant une ambiguïté surperflue. Ainsi, lorsque le complément indirect est mobile, c'est un complément du verbe, sinon c'est un complément du nom.

La figure 3 illustre le comportement double du complément prépositionnel dans le cas d'une relativisation[6].



Les réclamations que fait la LFP auprès du ministre des sports (sont justifiées)



Les réclamations auprès du ministre des sports que fait la LFP (sont justifiées)

FIG. 3 – Représentation de la relativisation du groupe prédicatif

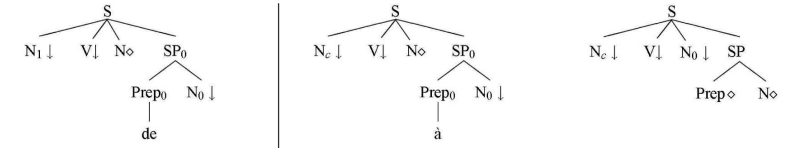
Le passif lexical et le causatif synthétiques ne sont pas simplement des variations de la morphologie active. En effet un changement de fonction est opéré :

- pour le passif lexical, l'inversion des arguments est constatable et l'effacement de l'agent est autorisé. Ce qui le rapproche du passif verbal.
  - Phrase de base : *L'ONU inflige des sanctions au programme nucléaire iranien.*
    Passif lexical : *Le programme nucléaire iranien essuie des sanctions de l'ONU.*
- pour le causatif synthétique, le sujet initial du verbe support devient ou objet indirect introduit par *à* (1) ou objet direct (2) et un nouvel argument (le causateur) est introduit[7].
  - Phrase de base : *La communauté a de la peine.*
    (1) Causatif synthétique : *Ces mauvais résultats font de la peine à la communauté.*
  - Phrase de base : *Les riziculteurs coréens sont en colère.*
    (2) Causatif synthétique : *L'ouverture du marché met les riziculteurs coréens en colère.*

On donne figure suivante quelques représentations dans la grammaire TAG du passif lexical et du causatif synthétique. Le trait *vconv* indique la forme du verbe converse sélectionné par le nom prédicatif, le trait *vcaus* indique la forme de l'opérateur causatif à utiliser.

[6]Noter que contrairement aux arbres des familles à verbes pleins, les arbres représentant la relativisation du groupe prédicatif sont des arbres initiaux et non des arbres auxiliaires. En effet, lorsque le nom prédicatif est l'antécédent de la relative, celui-ci demeure la tête lexicale.

[7]Les constructions à attribut de l'objet ayant un comportement très similaire, on les considère comme des variantes de ce phénomène : *Le ministre trouve les riziculteurs en colère.*



V :<vconv> = écoper | essuyer

Le nucléaire iranien essuie des sanctions de l'ONU

V:<vcaus> = causer | faire

Ces mauvais résultats font de la peine à la communauté

V:<vcaus> = rendre | mettre

L'ouverture du marché met les riziculteurs coréens en colère

FIG. 4 – Arbres représentant un passif lexical et des causatifs synthétiques

Outre, les quelques traits déjà indiqués, tous les arbres renseignent au niveau de leurs traits des fonctions explicites, qui permettent de noter des contraintes distributionnelles.

La sous-catégorisation réalisée étant de nature syntaxique, les fonctions déjà définies pour les verbes pleins, comme le sujet, l'objet, l'objet indirect (introduit par *à*, *de* ou une autre préposition), le complément locatif (de provenance ou non) et le complément d'agent[8] sont également utilisées dans les arbres des familles à verbe support[9]. D'autres fonctions sont néanmoins introduites comme les fonctions d'attribut prépositionnel et de complément du nom. Ces derniers peuvent être introduits par *à, de*[10] ou une autre préposition, ou être à valeur locative.

Les fonctions sous-catégorisées par un verbe sont susceptibles de se réaliser en position canonique ou en position extraite (relative, interrogative ou clivée). Les fonctions sous-catégorisées par un nom peuvent apparaître en position canonique ou sous la forme d'un déterminant possessif.

En somme, le traitement des verbes pleins et celui des verbes supports apparaît en partie similaire, malgré des différences représentationnelles évidentes. Nous allons maintenant examiner les modifications ou extensions à apporter à la représentation hiérarchique des verbes pleins pour y intégrer le traitement des constructions à verbe support.

[8](Candito 1999) définit également les fonctions d'infinitive et d'interrogative. Ces fonctions n'ont pas été reprises par (Abeillé 2002).

[9]Ainsi, le nom prédicatif qui est la tête sémantique de la construction à verbe support, porte une fonction par rapport à sa tête syntaxique (le verbe support).

[10]Nous suivons l'analyse proposée par (Godard, 1992) qui différencie les 1er et 2nd argument. Le premier argument a la particularité de pouvoir être réalisé comme un déterminant possessif, ce que ne peut faire le 2nd argument. Il y a deux moyens d'être premier argument : soit l'argument est premier d'origine, soit il le devient. En effet, un second argument peut être promu et alors accéder au statut de premier argument, si le premier argument d'origine est supprimé ou rétrogradé (c'est-à-dire s'il accède à la fonction de complément du nom introduit par la préposition *par*). Ainsi sont expliquées les possibilités d'apparition du déterminant possessif :

- *Son (= de Vermeer) portrait de la Laitière*
- **Son (= de La Laitière) portrait de Vermeer*
- *Son (= de la Laitière) portrait par Vermeer*
- *Son (= de Vermeer | = de la Laitière) portrait*

3 Intégration à la MG des verbes pleins

L'implémentation que nous avons réalisée utilise le compilateur de Métagrammaire défini par (Candito, 1999). Nous avons pour cela examiné la métagrammaire des verbes pleins et celle des constructions à verbe support pour en saisir toutes les particularités. Leurs différences majeures se situent principalement au niveau hiérarchique : la MG des verbes pleins réalise une hiérarchie très ordonnée, dont énormément de classes n'ont pour seul but que de regrouper un phénomène. A l'inverse, la hiérarchie des constructions à verbe support est beaucoup plus plate. Nous avons tout de même choisi de privilégier la hiérarchisation des verbes pleins, celle-ci étant historiquement la plus ancienne. D'autres différences existent, la plus importante se situant dans le fait que la hiérarchie des constructions à verbe support ne décrit des bouts d'arbres qu'en dimensions 2 et 3. Ce fait est pour nous important, c'est pourquoi nous le retenons également. Au final, la nouvelle hiérarchie tente de présenter l'information la plus pertinente possible, en sélectionnant les points les plus intéressants de chaque hiérarchie originelle.

3.1 Extension et modification des classes de dimension 1

Le rôle initial de la dimension 1 étant d'établir la sous-catégorisation de chaque famille, l'intégration des familles représentant les constructions à verbe support à la hiérarchie des verbes pleins ne présente pas de difficultés majeures.

Néanmoins, dans l'optique de créer une hiérarchie compacte dont les classes sont facilement réutilisables, il est préférable de revoir la façon dont l'information est renseignée. C'est pourquoi, on va prévoir dès la déclaration en dimension 1, la forme de réalisation de la tête syntaxique et de ses compléments ainsi que celui de la tête lexicale, ce qui évitera de le répéter dans les classes de dimension 2 et 3, et réduira au final le nombre total de classes décrites par ces dimensions. Autrement dit, en dimension 1, l'information ne sera plus seulement fonctionnelle, mais également catégorielle ; le type (nœud ancre ou nœud à substituer) de chaque variable mise en jeu sera en outre clairement renseigné. Les classes de dimensions 2 et 3 s'affranchiront ainsi en grande partie du formalisme TAG, ce rôle étant dès lors plus spécifiquement réservé à la dimension 1.

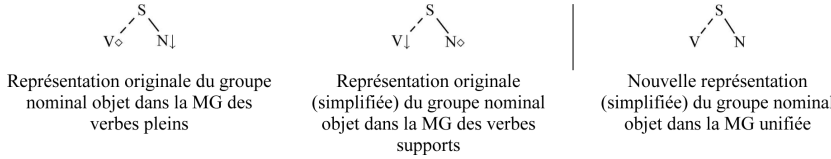


FIG. 5 – Un exemple du contenu de classes de dimension 3 avant et après modification

Les familles des verbes pleins déclarent donc en dimension 1, l'ancre verbale comme une tête syntaxique et ses arguments nominaux ou clitiques comme des nœuds à substituer. Les familles représentant les constructions à verbe support déclarent, quant à elles, le verbe support comme la tête syntaxique, ses arguments (excepté le nom prédicatif) comme des nœuds à substituer et le nom prédicatif comme une ancre lexicale.

Dimension 1	
Famille à verbe plein Contenu de la classe (n0Vn1)	Famille à verbe support Contenu de la classe (n0vN)
Var *cmpl0* = $N_0 \downarrow$ ou $Cl_0 \downarrow$ - Fonc. Sujet Var *cmpl1* = $N_1 \downarrow$ ou $Cl_1 \downarrow$ - Fonc. Objet Var *verbe* = V$\diamond$ - Tête syntaxique et lexicale	Var *cmpl0* = $N_0 \downarrow$ ou $Cl_0 \downarrow$ - Fonc. Sujet Var *cmpl1* = N$\diamond$ - Tête lexicale - Fonc. Objet Var *verbe* = V$\downarrow$ - Tête syntaxique

FIG. 6 – Un exemple de déclaration de variables en dimension 1

3.2 Restriction et extension des classes de dimension 2

C'est principalement la dimension 2 qui va demander le plus d'efforts de codage. Outre les nouvelles classes à ajouter permettant de renseigner le groupe nominal complexe, les possibilités de double analyse, les supports converses et les opérateurs causatifs, il va falloir empêcher leur croisement avec les classes des familles des verbes pleins par le biais de croisements contraints, en spécifiant que ces nouvelles réalisations ne concernent que les familles à ancre nominale.

3.2.1 Le groupe nominal complexe

Comme c'est le verbe support qui sous-catégorise les arguments en dimension 1, la formation du groupe nominal complexe n'est pour l'instant pas assurée. Il faut en fait établir une redistribution pour que cette formation soit possible. La sous-catégorisation va alors totalement concerner le nom prédicatif. Le groupe prédicatif qui portait initialement une fonction par rapport au verbe support va donc voir cette fonction supprimée et le nom prédicatif va alors présenter à la fois les emplois de tête syntaxique et de tête sémantique[11].

Comme le groupe nominal complexe peut avoir plusieurs sources possibles (la construction à verbe support active, la construction à verbe support passive et la construction converse), cette création aura lieu après les redistributions actives, passives et converses. Le sujet du verbe support (initial ou non) deviendra alors un complément du nom en *de* (premier argument), alors que ses autres compléments deviendront compléments du nom (second argument pour le complément prépositionnel introduit par *de*).

3.2.2 Le comportement double des compléments prépositionnels et phrastiques

Même si la grammaire TAG ne fournit finalement qu'une seule analyse pour une phrase particulière, le comportement double des compléments doit être matérialisé au niveau de la dimension 2. En effet, les compléments initialement renseignés comme des compléments du verbe support peuvent également apparaître comme des compléments du nom, notamment en cas d'extraction conjointe avec le nom prédicatif.

Le passage des fonctions verbales indirectes aux fonctions de compléments du nom se révèlera cependant insuffisant pour autoriser la formation d'un arbre, car aucune ossature ne sera dessinée pour y attacher ces compléments.

[11] Le fait que les structures des nominalisations soient dérivées des constructions à verbe support ne représente pas une prise de position théorique : la direction de la dérivation est imposée par le fait que les CVS représentent le centre d'intérêt de l'article.

La dimension 2 n'est en effet pas le strict pendant de la dimension 1 : elle fournit non seulement une sous-catégorisation finale, mais apporte également toute l'ossature sur laquelle viennent se placer les compléments à réaliser.

Il faudra donc réintroduire toute la morphologie verbale déjà renseignée, et la composer avec les changements de fonction annoncés.

3.2.3 Le passif lexical et le causatif synthétique

Enfin, les classes représentant le passif lexical et le causatif synthétique vont effectuer les changements de fonctions nécessaires.

Ces classes devront ensuite être composées avec les classes de la morphologie verbale déjà décrites pour les verbes pleins, pour permettre la description des constructions utilisables pour les familles à verbe support.

Tous ces changements ne sont que des ajouts à la hiérarchie des verbes pleins originelle, mais ils permettent de souligner le double travail effectué par la dimension 2 : il y a d'une part les changements de fonction à décrire, et d'autre part, une ossature à dessiner. En général, les deux opérations sont liées, mais cela n'est pas toujours le cas.

3.3 Extension et modification des classes de dimension 3

Il nous reste enfin à examiner les extensions et modifications à apporter aux classes de dimension 3. Le fait que les nœuds feuilles arguments aient été débarrassés de leur type et que cette information soit désormais renseignée par la dimension 1, permet de réduire l'importance de la dimension 3 en termes de classes.

3.3.1 Représentations des fonctions spécifiques aux familles à verbe support

Comme les familles à verbe support utilisent, en plus des fonctions déjà définies pour les verbes pleins, de nouvelles fonctions, il faut ajouter ces fonctions à la description originelle. Les classes ajoutées ne viennent pas gêner la représentation des verbes pleins puisque les fonctions ajoutées ne les concernent pas et que des contraintes sur les croisements ont été décrites pour que des redistributions inadéquates pour les verbes pleins ne soient pas réalisées.

3.3.2 Représentation des relatives

Le traitement en TAG des relatives pose un problème particulier. En effet, les arbres des relatives sont en général des arbres auxiliaires, dans lesquels l'élément relativisé constitue le nœud pied. Mais, dans le cas des verbes supports lorsque le groupe prédicatif est relativisé, les arbres des relatives sont des arbres initiaux dont le nom prédicatif constitue l'ancre. Le fait d'avoir déclaré explicitement en dimension 1 le type des compléments du verbe va donc produire un conflit de valeur pour les compléments à substituer (la valeur initialement apportée par la dimension 1 n'indiquant pas un nœud pied).

Il faut donc introduire au sein de la classe représentant les relatives une règle de révision spécifiant qu'un nœud initialement à substitution doit devenir un nœud pied[12].

3.3.3 « Descente » du nom prédicatif

Originellement, les classes représentant les compléments nominaux dans la hiérarchie des verbes pleins, reliaient directement l'argument nominal à sa racine. Or, pour pouvoir prendre en compte le fait que le nom prédicatif peut avoir un complément du nom, il est nécessaire de modifier cette description. En effet, le nom prédicatif n'est dans ce cas plus directement relié à sa racine.

Pour traiter la « descente » du nom prédicatif dans un syntagme, il va donc falloir redessiner l'arbre syntagmatique. On va alors avoir recours à une relation sous-spécifiée entre un N (représentant potentiellement un groupe nominal avec ses compléments) portant une fonction, et un nom qui n'en porte pas. Le N portant la fonction syntaxique va quant à lui être directement relié à sa racine. Ainsi, lorsqu'un syntagme prépositionnel est complément du nom, ce groupe se place à droite du N ne portant pas de fonction. En revanche, lorsque aucun élément ne vient s'insérer, le lien est ramené à 0 : le N portant la fonction et le N n'en portant pas sont alors identifiés comme un seul nœud. Ce qui permet de représenter les arbres adéquats en faisant porter la fonction syntaxique au groupe prédicatif et non à sa tête, comme l'illustre la figure 7[13].

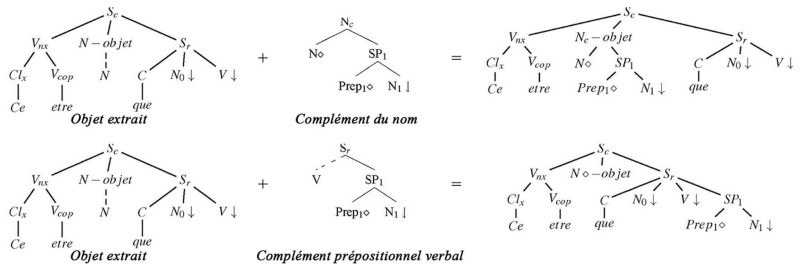


FIG. 7 – Exemple de « descente » du nom prédicatif avec l'objet clivé

4 Conclusion

Nous avons présenté dans cet article les caractéristiques de la grammaire TAG des verbes supports du français. Notre but était double : nous voulions d'une part mettre en valeur les éléments qui la distinguent de la grammaire TAG des verbes pleins, et d'autre part insister sur les modifications à apporter à la métagrammaire des verbes pleins originelle, pour y intégrer ces nouvelles descriptions.

[12] Il reste donc ici une information TAG que l'on ne peut pas supprimer, sauf à intervenir sur les croisements en créant cette règle de révision hors des classes de dimension 3 et en la faisant s'appliquer par le biais d'un croisement contraint.

[13] Les arbres représentés par la figure 7 ne sont pas le reflet exact du contenu des classes de dimension 3, mais les résultats intermédiaires de la formation des arbres lors du croisement des classes.

L'implémentation de ces modifications est achevée, et nous n'avons pas rencontré de difficulté rédhibitoire concernant l'écriture de la hiérarchie. C'est principalement le nommage des variables qui a été le plus difficile à mettre en œuvre. En effet, la hiérarchie des verbes pleins utilise des noms de variables sans rapport direct avec leur contenu. Il a bien souvent fallu parcourir l'ensemble de la hiérarchie pour identifier exactement le rôle d'une variable pour l'adapter ensuite aux constructions à verbe support. Quoiqu'il en soit, nous avons fait en sorte que les variables soient désormais clairement identifiées lorsqu'elles sont utilisées, et la maintenance future de cette hiérarchie devrait donc se révéler plus aisée.

Mise à part la dimension 1, les autres dimensions sont maintenant beaucoup plus neutres linguistiquement puisqu'elles ont éliminé une grande partie de l'information TAG (dans le sens où la nature d'un nœud n'est plus renseignée dans ces dimensions). 59 classes ont été créées, près de 160 ont été modifiées. Au final, la nouvelle hiérarchie compte plus de 330 classes, ce qui représente une augmentation de son volume d'un peu plus de 20%. En termes d'arbres, la hiérarchie produit près de 8000 arbres pour les verbes pleins, et près de 10000 arbres pour les constructions à verbe support.

L'organisation de la métagrammaire étant modulaire, les divers changements intervenus dans les classes n'ont pas eu d'incidence sur le reste de la grammaire. Aucune contradiction dans le reste de la ressource n'a été enregistrée, et la grammaire résultante a finalement une couverture aussi importante que la grammaire d'origine.

Références

ABEILLÉ A. (1991). *Une grammaire lexicalisée d'arbres adjoints pour le français*. PhD thesis, Université Paris 7.

ABEILLÉ A. (2002). *Une grammaire électronique du français*. Paris : CNRS Editions.

A. ABEILLÉ & O. RAMBOW, Eds. (2000). *Tree Adjoining Grammars*. Stanford : CSLI Publications.

BARRIER S. (2006). *Une métagrammaire pour les noms prédicatifs du français : développement et expérimentations pour les grammaires TAG*. PhD thesis, Université Paris 7.

CANDITO M.-H. (1999). *Une métagrammaire pour les noms prédicatifs du français*. PhD thesis, Université Paris 7.

CRABBÉ B. (2005). *Représentation informatique de grammaires fortement lexicalisées*. PhD thesis, Université Nancy 2.

GODARD D. (1992). Extraction out of NP in French. *Natural Language and Liguistic Theory*, **10**.

Un nouveau cadre de factorisation pour les grammaires d'arbres adjoints

Nicolas BARRIER
Laboratoire Lattice - Université Paris VII
30, rue du château des rentiers, 75013 Paris
nbarrier@linguist.jussieu.fr

Résumé. Dans cet article, nous présentons un nouveau cadre de factorisation pour les grammaires d'arbres adjoints qui permet de combiner automatiquement l'information grammaticale. Ce cadre facilite l'expression des contraintes permettant la combinaison des différents fragments arborescents correspondant chacun à la réalisation syntaxique d'un phénomène linguistique, et permet une meilleure réutilisation des descriptions partielles utilisées. L'invariant syntaxique est utilisé comme structure de contrôle.

Mots-clés : MG, génération, sous-spécification, héritage, croisement, unification, TAG.

Keywords: MG, generation, underspecification, inheritance, crossing, unification.

1 Introduction

Depuis leur apparition dans le milieu des années 90, les méta-grammaires (dorénavant MG) ont facilité la réalisation de nombreux projets linguistiques. Pour les grammaires dans lesquelles ce mode de représentation a été adopté, elles ont permis d'étendre de façon significative la couverture syntaxique.[1] Un arbre TAG élémentaire peut en effet être considéré comme la combinaison de plusieurs fragments arborescents,[2] dont les croisements correspondant à la réalisation syntaxique de plusieurs phénomènes indépendants ont été systématisés.[3] La façon dont ces unités syntaxiques d'ordre supérieur sont combinées entre elles constitue de fait la problématique générale des méta-grammaires.

[1] Le formalisme des grammaires d'arbres adjoints oblige à spécifier dans l'arbre élémentaire d'un verbe, d'un nom ou d'un adjectif, la position de ses arguments syntaxiques. Une grammaire TAG comporte donc en pratique un nombre conséquent d'arbres élémentaires dont certains éléments sont répétés à l'identique dans d'autres structures sans aucun mécanisme de partage.

[2] On parle de combinaison pour désigner un assemblage d'éléments dans un arrangement déterminé.

[3] Les données de la figure 1 font apparaître des structures arborescentes. Il s'agit en fait d'un raccourci. (Vijay-Shanker & Schabes, 1992) ont proposé d'employer un langage formel permettant de factoriser les parties communes à certaines réalisations dans un graphe d'héritage unique. Ils utilisent pour cela un langage de type prédicatif, sans quantification, dont le connecteur privilégié est la conjonction. Les termes du langage sont des constantes, reliées entre elles par les quatre opérateurs binaires suivants : la dominance large (notée $\lhd*$), la parenté ($\lhd$), la précédence ($\prec$), et l'égalité ($\approx$). Ces constantes sont augmentées d'une structure de traits qui permet de décrire leur contenu (catégorie du nœud et qualité). La dominance large est représentée graphiquement par un trait pointillé, alors que la parenté est notée d'un trait plein. L'arbre droit de la figure 1 correspond en revanche à un arbre TAG classique avec des nœuds à substituer (notés $\downarrow$) et une ancre (notée $\diamond$) (Joshi & Schabes, 1997).

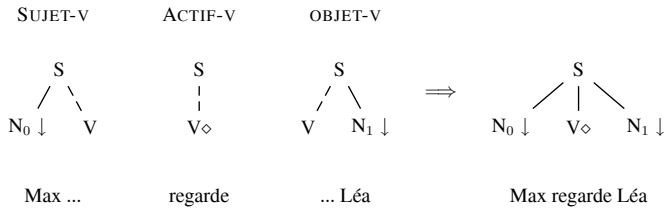


FIG. 1: Un exemple de conjonction de descriptions partielles

Jusqu'ici, on avait principalement distingué deux approches concurrentes pour combiner l'information grammaticale : l'une mise au point par (Crabbé, 2005) et l'autre développée par (Candito, 1996).[4] La différence entre ces deux implémentations réside dans l'utilisation d'un langage de contrôle pour le premier système, alors que le second repose sur l'utilisation d'un algorithme de croisement qui fixe et impose les domaines d'informations retenus dans les descriptions.

Le cadre formel défini par (Crabbé, 2005) se caractérise donc par un aspect déclaratif accru. Il permet d'exprimer de façon contrainte la manière dont les descriptions arborescentes sont combinées entre elles. Il repose sur l'utilisation d'un langage de contrôle, qui par l'intermédiaire des connecteurs logiques *ou* et *et*, permet une reconnaissance explicite des alternatives de diathèses[5] et des alternatives de réalisations.[6]

Les structures élémentaires ainsi regroupées sont alors combinées entre elles de façon à produire un arbre élémentaire qui soit en rapport avec la sous-catégorisation finale d'un lemme[7] pour une diathèse donnée, selon les indications fournies par l'utilisateur. On peut ainsi exprimer directement des alternatives de sous-catégorisation, conjointement aux alternatives de réalisations définies pour une fonction syntaxique :

$$\textit{Verbe-transitif} \quad \rightarrow$$
$$(\ \textit{Sujet} \ \wedge \ \textit{Actif} \ \wedge \ \textit{Objet}\ ) \ \vee \ (\ \textit{Sujet} \ \wedge \ \textit{Passif} \ \wedge \ \textit{Par-objet}\ )$$

En revanche, il n'est pas possible dans un tel système d'exprimer facilement les contraintes de combinaison qui interagissent entre classes : le système est verbeux et ne permet pas réellement de partager l'information syntaxique en dehors du cadre des descriptions partielles. L'utilisateur liste donc de façon exhaustive les croisements qu'il souhaite réaliser.

Chez (Candito, 1996) au contraire, cette sous-catégorisation est obtenue par calcul computationnel après croisement des informations syntaxiques issues des classes de dimension 1 et de

[4]La liste des compilateurs implémentés ne se limitent pas aux deux implémentations mentionnées. On trouve également dans la littérature de nombreux autres projets linguistiques ou informatiques ayant tous pour but de factoriser l'information grammaticale. (Gaiffe *et al.*, 2002) ont proposé par exemple d'utiliser un système reposant sur l'utilisation d'un mécanisme de *besoins* et de *ressources*, alors que d'autres au contraire ont préféré implémenter un système à base de règles lexicales (Becker, 2000).

[5]On parle de diathèse pour désigner la voix associée au verbe et à son auxiliaire (active, passive, réfléchie, ou moyenne). Chacune des voix se manifeste par des flexions verbales spécifiques : désinences ou préfixes, formes différentes des auxiliaires, etc.

[6]On parle d'une alternative de réalisations pour désigner par exemple l'alternance sujet nominal - sujet clitique.

[7]Valence, catégorie syntagmatique des arguments, fonction syntaxique, etc.

dimension 2.[8] Les arguments syntaxiques qui sont en rapport avec cette sous-catégorisation sont ensuite réalisés selon les indications fournies en dimension 3. Le calcul d'un arbre élémentaire repose donc sur l'utilisation d'un algorithme de croisement, qui guide le processus général de dérivation, mais dont on ne peut s'affranchir en pratique.

Cet algorithme de croisement suppose par ailleurs que les classes croisées sont obtenues par héritage multiple.[9] Or, sur le plan pratique, rien ne permet de justifier une telle décision. Cela conduit en effet à des situations aberrantes, dans lesquelles il n'est pas possible de partager efficacement l'information grammaticale. Ce comportement s'observe particulièrement en anglais dans les cas de réalisations bi-transitives, où l'argument bénéficiaire du verbe apparaît conjointement avec un objet nominal.

(1) a. The Commission sent [the Governor] [a progress report]
b. He gave [the audience] [a fascinating glimpse]

Pour rendre compte de ces constructions dans les implémentations proposées, il faut alors distinguer deux fonctions syntaxiques différentes pour des arguments qui se réaliseront sur le plan syntagmatique sans aucun mécanisme de partage. L'héritage multiple ne permet pas en effet d'hériter plusieurs fois directement d'une même classe ; l'information grammaticale ne peut donc jamais être dupliquée.

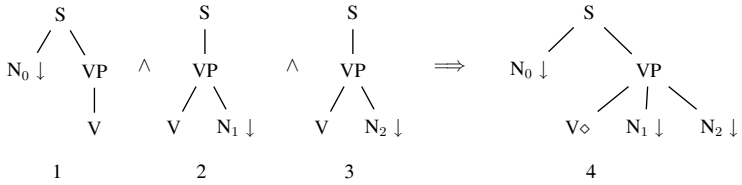


FIG. 2: Les structures 2 et 3 ne partagent aucune information syntaxique

L'héritage est donc utilisé chez (Candito, 1996) de façon contradictoire pour satisfaire deux besoins différents : celui de la *classification* d'une part, et celui de la *construction*. Les variables introduites dans les descriptions partielles ont pour elle une portée globale, et doivent s'unifier entre elles si elles portent le même nom. Or, ceci est problématique d'un point de vue pratique. Une même description peut être réutilisée plusieurs fois dans une même construction. L'utilisateur est donc amené dans cette implémentation à gérer lui-même l'espace de nom qui est associé aux variables, ce qui limite en pratique le partage de l'information syntaxique et se révèle rapidement ingérable sur des exemples non triviaux.

[8] Un arbre élémentaire est défini dans la MG comme une combinaison contrôlée de plusieurs domaines d'informations appelés dimensions. Le terme lui-même est issu des travaux de (Koenig & Jurafsky, 1994) sur la représentation du lexique. Les classes de dimension 1 définissent une sous-catégorisation initiale établie en termes fonctionnels, alors que les classes de dimensions 2 et 3 correspondent respectivement aux alternatives de diathèses et aux alternatives de réalisations.

[9] Les classes croisées désignent dans une méta-grammaire les classes obtenues par croisement automatique de plusieurs phénomènes linguistiques lors de la compilation d'une MG en une TAG. L'ensemble des classes croisées est caractérisé par des principes de bonne formation, dont une partie est imposée par le formalisme cible et non par la langue.

2 Un premier niveau d'abstraction

Pour solutionner les problèmes évoqués précédemment, nous partons dans un premier temps d'un système descriptif permettant l'expression des alternatives de réalisations pour un argument syntaxique donné. Les descriptions partielles que nous utilisons sont combinées entre elles par l'intermédiaire d'un langage de contrôle qui repose sur l'utilisation des connecteurs logiques *ou* et *et*.

Pour résoudre le problème de la *réutilisabilité* évoqué en section 1, nous proposons d'augmenter les descriptions partielles d'un ensemble de traits regroupés dans une nouvelle structure que nous appellerons *hypertag*.[10] Cette structure décrit de façon simple le contenu d'une classe, et permet l'expression de généralisations en paramétrant le type de certains constituants (*template*).[11] Pour en illustrer le fonctionnement, nous reprenons l'exemple des constructions bi-transitives pour l'anglais. Nous associons à l'index des arguments nominaux, un trait NP $\boxed{x}$ dans la structure de traits. Cette structure indique par ailleurs que la description partielle qui est associée à l'hypertag réalise un complément verbal de nature nominale.

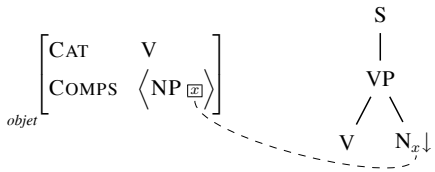


FIG. 3: L'information syntaxique est partagée conjointement entre deux structures

Pour pouvoir générer un arbre élémentaire qui soit en rapport avec cet argument, il faut ensuite instancier la description partielle définie en figure 3. Les classes croisées ne sont alors plus obtenues par héritage, mais par composition. Un arbre élémentaire apparaît donc dans la méta-grammaire comme un objet composite, dont les différents champs sont des instances de certaines réalisations.[12]

Ce même mécanisme peut encore être utilisé pour partager l'information grammaticale entre compléments prépositionnels. On peut ainsi spécifier la valeur lexicale d'une préposition, et partager localement l'indice d'un nœud avec un autre. On n'a donc pas besoin comme chez Candito de dupliquer les descriptions partielles selon la nature de la préposition.

En pratique, on peut donc réduire de façon significative la taille d'une méta-grammaire, tout en lui assurant une meilleure lisibilité.[13] On minimise ainsi les risques d'erreur tout en facilitant la maintenance et l'extension de la grammaire.

[10]Nous appelons donc hypertag une matrice de traits associée à une classe ou à un ensemble de classes, ce qui en définitive s'éloigne de la définition fournie par (Kinyon, 2000).

[11]Ces patrons syntaxiques vont nous permettre de réduire le nombre d'erreurs dues à la recopie de certaines classes par copier-coller (copier-coller assassin).

[12]On peut ainsi réutiliser plusieurs fois une même description, sans avoir à dupliquer par ailleurs certaines réalisations (classes) qui ne diffèrent que par leurs constantes (et non par leur contenu). En conséquence, le problème de la portée des variables ne se pose pas réellement, ou du moins pas directement, dans la mesure où les desriptions partielles sont instanciées au cas par cas. Il n'y a donc plus de conflit de variables lors de la génération d'une classe croisée, puisque par définition une instance d'une classe est unique.

[13]Pour rendre compte des différents compléments prépositionnels apparaissant dans un arbre élémentaire ancré par un verbe, il fallait chez Candito plus de 20 classes différentes. Dans notre implémentation, 8 classes suffisent.

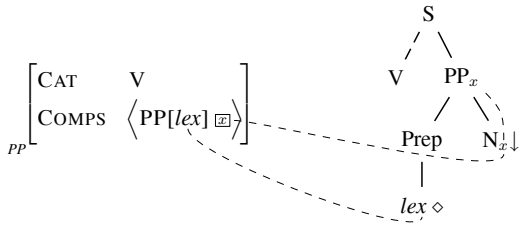


FIG. 4: Les templates peuvent également intervenir sur des items lexicaux

3 L'invariant syntaxique comme structure de contrôle

Pour générer un arbre élémentaire, nous abandonnons maintenant notre langage de contrôle, pour une structure de traits typés qui guide le processus général de dérivation. Au lieu de lister de façon exhaustive les croisements à réaliser pour une diathèse et une sous-catégorisation données, nous introduisons une nouvelle structure qui par *unification* avec les hypertags définis précédemment produira un arbre élémentaire en rapport avec cette sous-catégorisation.

On fournit ci-après un exemple de réalisation pour la sous-catégorisation d'un verbe comme *parler*. Cette structure identifie de façon simple les arguments syntaxiques à réaliser pour un prédicat, et rappelle que ceux-ci sont sous la gouvernance d'un verbe.[14]

$$\begin{bmatrix} \text{VALENCE} & \begin{bmatrix} \text{CAT} & \text{V} \\ \text{SUJ} & \langle \text{NP } \boxed{0} \rangle \\ \text{COMPS} & \langle \text{PP}[\grave{a}]\ \boxed{1}, \text{PP}[de]\ \boxed{2} \rangle \end{bmatrix} \\ \text{DIATHÈSE} & \text{Actif} \end{bmatrix}$$

FIG. 5: Sous-catégorisation d'un verbe comme *parler*

Sont résumées dans cette structure les principales caractéristiques d'un verbe transitif sous-catégorisant deux compléments de nature prépositionnelle dans le cadre d'une construction personnelle active. La structure de la figure 5 doit alors s'unifier aux hypertags définis dans la hiérarchie, pour produire par combinaison un arbre élémentaire qui soit en rapport avec les indications fournies dans la figure. A la création de la classe croisée, on dispose donc de quatre instances différentes, correspondant chacune à la réalisation d'une classe donnée (cf. figure 8).[15]

Malheureusement, telles qu'elles sont exprimées, ces contraintes ne permettent pas encore de partager efficacement l'information grammaticale. Formellement, nous sommes encore obli-

[14] On ne réalise pas par exemple de la même façon les arguments qui sont sous la gouvernance d'un verbe ou sous la gouvernance d'un adjectif.

[15] Pour s'assurer de la bonne formation des arbres élémentaires produits, on introduit dans le compilateur quelques contraintes de bonne formation, qui s'apparentent aux principes définis en LFG (Bresnan, 1982) (unicité, complétude et cohérence). On dispose donc en pratique d'un mécanisme pour rejeter des structures incorrectes d'un point de vue fonctionnel.

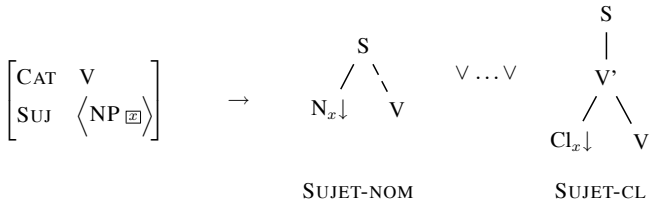


FIG. 6: Réalisation d'un sujet de nature nominale

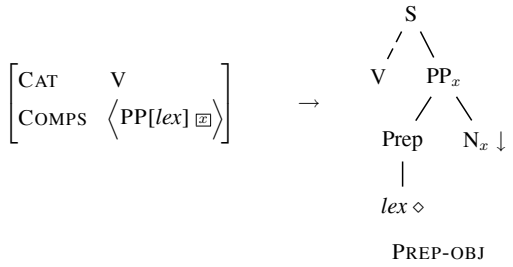


FIG. 7: Complément prépositionnel en position canonique

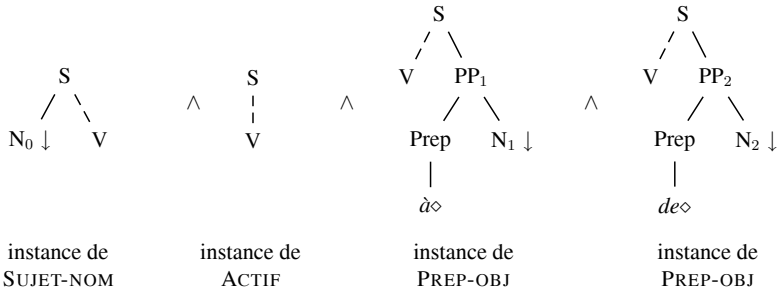


FIG. 8: Une conjonction de descriptions partielles compatible avec les informations syntaxiques présentées en figure 5

gés comme chez (Crabbé, 2005) de lister chacune des réalisations attendues pour une sous-catégorisation et une diathèse données. L'idée est donc d'obtenir ces informations par l'intermédiaire du lexique, en sous-spécifiant certains arguments. Nous introduisons pour cela un trait HEAD à valeur lexicale, qui va nous permettre de rendre compte de la sous-catégorisation initiale d'un prédicat. Ce trait est largement sous-spécifié et facilite la mise en correspondance du lexique avec la grammaire. C'est par son intermédiaire que l'on obtiendra pour la syntaxe les informations nécessaires à la saturation du trait VALENCE. Un verbe comme *parler* sera donc associé à la structure de la figure 9.[16]

$$\begin{bmatrix} \textsc{Head} & \begin{bmatrix} \textsc{Cat} & \text{V} \\ \textsc{Arg-st} & \left\langle \boxed{0}\,\text{NP}, \boxed{1}\,\text{PP}[\grave{a}], \boxed{2}\,\text{PP}[de] \right\rangle \end{bmatrix} \end{bmatrix}$$

FIG. 9: Ressource lexicale associée au verbe *parler*

Un verbe comme *dormir* disposera en revanche d'un unique argument. Or, pour rendre compte de ces deux constructions, il faut pouvoir exprimer des contraintes qui s'appliquent indifféremment aux deux types de verbes. La structure de la figure 5 doit donc être révisée pour s'adapter aux spécificités du lexique. On doit donc pouvoir spécifier qu'un verbe qui se réalise à l'actif attend au moins la réalisation d'un argument syntaxique qui apparaît dans la valence du verbe comme un sujet – les arguments restants devant se réaliser comme des compléments du verbe (cf. liste $\boxed{1}$ de la figure 10).[17]

$$\begin{bmatrix} \textsc{Head} & \begin{bmatrix} \textsc{Cat} & \text{V} \\ \textsc{Arg-st} & \left\langle \boxed{0}\,\text{NP} \right\rangle \oplus \boxed{1}\,\text{list} \end{bmatrix} \\ \textsc{Valence} & \begin{bmatrix} \textsc{Cat} & \text{V} \\ \textsc{Suj} & \left\langle \text{NP}\,\boxed{0} \right\rangle \\ \textsc{Comps} & \boxed{1} \end{bmatrix} \\ \textsc{Diathèse} & \text{Actif} \end{bmatrix}$$

FIG. 10: Contraintes de combinaison opérant pour un verbe à l'actif

Pour rendre compte des tournures passives, il nous faudra par contre mettre à jour les informations issues de la structure argumentale. On utilisera pour cela les indices référentiels définis entre compléments. On identifiera ainsi le premier argument $\boxed{0}$ du trait HEAD au complément prépositionnel en *par* du trait VALENCE. L'argument $\boxed{1}$ de la structure argumentale passe ensuite en position sujet (cf. figure 11).

Telles quelles bien sûr ces données ne seront pas suffisantes pour rendre compte efficacement des contraintes qui pèsent sur la réalisation des verbes au passif, et il faudra ajouter des informa-

[16] Il est possible de noter des contraintes d'ordre sémantique dans ces structures par l'intermédiaire d'un trait CONT qui peut interagir avec le niveau syntaxique.

[17] L'opérateur $\oplus$ correspond à l'opérateur de concaténation de listes.

$$\begin{bmatrix} \textsc{Head} & \begin{bmatrix} \textsc{Cat} & \text{V} \\ \textsc{Arg-st} & \left\langle \boxed{0}\ \text{NP}, \boxed{1}\ \text{NP} \right\rangle \oplus \boxed{2}\ \text{list} \end{bmatrix} \\ \textsc{Valence} & \begin{bmatrix} \textsc{Cat} & \text{V} \\ \textsc{Suj} & \left\langle \text{NP}\ \boxed{1} \right\rangle \\ \textsc{Comps} & \left\langle \text{PP}[\textit{par}]\ \boxed{0} \right\rangle \oplus \boxed{2} \end{bmatrix} \\ \textsc{Cont} & \begin{bmatrix} \boxed{0} & \text{Agent} \\ \boxed{1} & \text{Patient} \end{bmatrix} \\ \textsc{Diathèse} & \text{Passif} \end{bmatrix}$$

FIG. 11: Contraintes de combinaison opérant pour un verbe au passif

tions supplémentaires concernant par exemple la fonction sémantique des arguments distingués (trait CONT).

*(2) a. Ce paragraphe comporte 15 lignes / * 15 lignes sont comportées par ce paragraphe*

*b. Paul a quitté sa ville natale / * Sa ville natale a été quittée par Paul*

*c. Cette jeune-fille respire la santé / * La santé est respirée par cette jeune-fille*

Les arguments d'un verbe se réalisent donc de façon différente selon la *diathèse* utilisée. Il faut spécifier la sous-catégorisation effective d'un prédicat en même temps que sa diathèse, ce qui revient en pratique à proposer un mécanisme déclaratif de redistribution.

4 Un autre exemple de factorisation

Dans cette section, nous montrons comment utiliser efficacement la structure que nous avons présentée en section 3 pour rendre compte des constructions attributives du français. Nous souhaitons pour cela réutiliser un sous-ensemble des descriptions partielles déjà définies pour les verbes, pour ne pas avoir à réimplémenter certains phénomènes d'accord ou de sous-catégorisation déja encodés.

Depuis (Stowell, 1978), on considère que les constructions à attribut du sujet sont issues de tournures prédicatives, où le verbe *copule* se présente comme un verbe à *montée*, qui sélectionne sur le plan sémantique un argument unique, celui de la petite proposition (*SC*).

(3) a. [$_P$ être [$_{SC}$ SN Adj]]

b. [$_P$ être [$_{SC}$ Jean intelligent]]

Afin de satisfaire à la contrainte voulant que toute forme verbale ait un sujet exprimé, le sujet de la petite proposition monte en position pré-verbale.

(4) a. [$_P$ [$_{NP}$ Jean] être [$_{SC}$ intelligent]]

Les verbes à montée du sujet sous-catégorisent donc deux arguments, mais ne sélectionnent que le complément prédicatif représenté par l'adjectif. Le sujet du verbe copule est en correspondance avec le sujet logique de l'attribut, mais le verbe copule ne sélectionne pas sémantiquement son sujet ; c'est le complément prédicatif qui assure ce rôle.

Nous résumons l'ensemble de ces observations dans la structure invariante de la figure 12. Le sujet logique de l'attribut y est représenté comme un élément qui se réalise sur le plan syntaxique comme le sujet d'un verbe d'une part et comme le sujet d'un adjectif de l'autre. Le premier propose un accord en personne et en nombre avec le verbe copule comme il le ferait pour les verbes pleins, alors que le second marque l'accord en genre et en nombre avec l'attribut. L'élément [0] NP de la structure argumentale apparaît donc à la fois comme le sujet du verbe copule et le sujet de l'adjectif. Les compléments de l'adjectif restent par contre sous sa gouvernance.[18]

$$\begin{bmatrix} \textsc{Head} & \begin{bmatrix} \textsc{Cat} & \text{A} \\ \textsc{Arg-st} & \langle [0]\,\text{NP} \rangle \oplus [1]\,\text{list} \end{bmatrix} \\ \textsc{Valence} & \begin{bmatrix} \textsc{Cat} & \text{V} \\ \textsc{Suj} & \langle \text{NP}\ [0] \rangle \\ \textsc{Comps} & \left\langle \text{AP} \begin{bmatrix} \textsc{Cat} & \text{A} \\ \textsc{Suj} & \langle \text{NP}\ [0] \rangle \\ \textsc{Comps} & [1] \end{bmatrix} \right\rangle \end{bmatrix} \\ \textsc{Diathèse} & \text{Actif} \end{bmatrix}$$

FIG. 12: Contraintes portant sur la réalisation d'un adjectif attribut du sujet

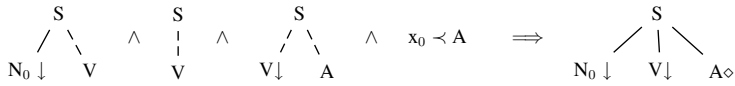


FIG. 13: Exemple de réalisation canonique d'une construction à attribut du sujet

Les constructions à attribut de l'objet reçoivent une analyse similaire, si ce n'est que l'accord est marqué entre l'objet du verbe et le sujet de l'adjectif.

Conclusion

Une méta-grammaire fait appel à plusieurs problématiques. Dans cet article, nous avons abordé les thèmes de la réutilisation des classes et celui de la génération des classes croisées. Nous

[18] Ces mêmes compléments peuvent apparaître également dans la structure invariante d'un adjectif épithète, mais tous ne sont pas compatibles avec ces constructions.

avons présenté un nouveau cadre de factorisation pour les grammaires d'arbres adjoints, qui se veut monotone et déclaratif. Ce cadre repose explicitement sur les principes de l'unification et permet de guider le processus général de dérivation en combinant automatiquement l'information grammaticale. Il faut pour cela ne pas se limiter à une simple information catégorielle sur les nœuds, et leur ajouter par ailleurs plusieurs contraintes (des indices, des fonctions, des couleurs, etc.).[19]

La structure que nous avons définie sépare donc clairement les contraintes de combinaison opérant entre classes et les descriptions partielles réparties dans la hiérarchie. Cette nouvelle structure simplifie la mise en correspondance du lexique avec la grammaire, et facilite également la réalisation de l'interface syntaxe/sémantique. Nous avons par ailleurs rendu explicite la notion de gouverneur syntaxique, qui nous permet de limiter la réalisation de certains arguments à certains prédicats. Enfin, nous sommes revenu plus succinctement sur le statut théorique d'une classe croisée, et avons montré qu'un arbre élémentaire apparaissait dans la méta-grammaire comme un objet composite dont les différents champs sont des instances de certaines réalisations.

Le cadre formel que nous avons présenté nous a permis en pratique de réduire de façon significative la taille d'une méta-grammaire, tout en lui assurant une meilleure lisibilité. Nous avons également pu réutiliser une partie des fragments arborescents déjà définis pour les verbes pour réaliser l'encodage des constructions attributives du français, ce qui n'avait jamais été réalisé auparavant (hors cas triviaux).

Références

BECKER T. (2000). *Patterns in metarules for TAG*, chapter 14. Tree Adjoining Grammars, formalisms, linguistic analysis and processing. CSLI Publications.

BRESNAN J. (1982). *The Mental Representation of Grammatical Relations*. Massachusetts : MIT Press : Cambridge.

CANDITO M.-H. (1996). A principle-based hierarchical representation of LTAGs. In *Proceedings of COLING-96*, Copenhagen, Denmark.

CRABBÉ B. (2005). *Représentation informatique de grammaires fortement lexicalisées - application à la grammaire d'arbres adjoints*. PhD thesis, Université de Nancy 2, Nancy.

GAIFFE B., CRABBÉ B. & ROUSSALANY A. (2002). A new metagrammar compiler. In *Proceedings of TAG+6*, Venice, Italy.

JOSHI A. & SCHABES Y. (1997). Tree-adjoining grammars. In G. ROZENBERG & A. SALOMAA, Eds., *Handbook of Formal Languages*, volume 3, p. 69–124. New York : Springer.

KINYON A. (2000). Hypertags. In *Proceedings of the 18th conference on Computational linguistics*, Morristown, NJ, USA : Association for Computational Linguistics.

KOENIG J.-P. & JURAFSKY D. (1994). Type underspecification and on-line type construction in the lexicon. In *Actes de la West Coast Conference in Linguistics 13*.

STOWELL T. (1978). What was there before *there* was here ? In *Proceedings of the 14th Regional Meeting of the Chicago Linguistic Society*, Chicago.

VIJAY-SHANKER K. & SCHABES Y. (1992). Structure sharing in lexicalized tree adjoining grammars. In *COLING 92*.

[19] Faute de place, nous ne pouvons aborder ici la problématique de la conjonction des descriptions partielles.

Semantic pregroup grammars handle long distance dependencies in French

Anne PRELLER
Lirmm–CNRS, 161, rue Ada, 343924 Montpellier Cedex 5
preller@lirmm.fr

Résumé. Nous présentons une grammaire de prégroupe traitant l'accord entre le sujet ou l'objet antéposé avec le participe passé, actif ou passif, comprenant les verbes modaux. La grammaire est munie d'une interprétation sémantique respectant les dépendances non-bornées.

Abstract. A pregroup grammar is presented which handles distant agreement of features in French, including modal verbs, clitics, relative pronouns, the compound past and the passive mode. The grammar has a semantic interpretation into predicate logic which captures the unbounded dependencies.

Mots-clés : grammaires catégorielles, grammaires de prégroupe, dépendances non-bornées, clitiques, pronoms relatifs, interprétation sémantique.

Keywords: categorial grammars, pregroup grammars, distant dependencies, agreement of features, French clitics, French pronouns, semantic interpretation.

1 Interpretation in predicate logic

Pregroup grammars belong to the family of categorial grammars and were introduced in (Lambek, 1999) as a simplification of the earlier syntactic calculus, now known as Lambek Calculus. Though categorial grammars based on Lambek calculus can be translated into pregroup grammars, the translated pregroup grammar may be stronger than the original one and therefore overgenerate. Moreover, the inherent higher order semantical interpretation of categorial grammars is lost for pregroup grammars. Here, we want to show that the meaning of a sentence with long distance agreement based on an analysis by pregroup grammars can be defined in two-sorted predicate logic. The semantical interpretation used here was introduced in (Preller, 2007b). The main idea is to accompany the lexical entries in the dictionary by one or more logical expressions translating the entry. The translation of a sentence is computed from the translation of the words and from a reduction to the sentence type.

A pregroup grammar consists of a dictionary, associating to each word a finite number of *types*. Types are strings of *simple types*, i.e. of the form

$$a_1^{(z_1)} \dots a_k^{(z_k)},$$

where $a_1, \dots, a_k$ are basic types and $z_1, ..., z_k \in \mathbb{Z}$. The set of *basic types* B is partially ordered by $\rightarrow$ and includes the syntactical types, e.g. the sentence type s. When parsing a sentence with

a pregroup grammar one assigns to each word a type from the dictionary and constructs a derivation to the sentence type $\boldsymbol{s}$ by the following rules

$$\begin{array}{ll} \text{(Induced step)} & Xa^{(z)}Y \to Xb^{(z)}Y \\ \text{(Generalized contraction)} & Xa^{(z)}b^{(z+1)}Y \to XY \end{array} ,$$

where X and Y are arbitrary types, a and b are basic and either z is even and $a \to b$ or z is odd and $b \to a$. In the following, we write 1 for the empty string a^ℓ for $a^{(-1)}$and a^r for $a^{(1)}$ and refer to them a *adjoints* of a, whereas $a^{\ell\ell} = a^{(-2)}$, $a^{rr} = a^{(2)}$ are *iterated* adjoints.

For example, consider the dictionary

$$\begin{array}{ll} \textit{Marie} & : \pi_{3\mathtt{fs}}, o \\ \textit{Jean} & : \pi_{3\mathtt{ms}}, o \\ \textit{examine} & : \pi^r_{3\mathtt{s}} \boldsymbol{s} o^\ell \end{array} .$$

The basic type $\pi_{3\mathtt{fs}}$ corresponds to 'subject third person feminine singular', or more generally, π_{pgn} to 'subject of person p, gender g and number n', where $p \in \{1,2,3\}$, $g \in \{\mathtt{m},\mathtt{f}\}$ and $n \in \{\mathtt{s},\mathtt{p}\}$. Here, $\mathtt{m}$ stands for 'masculine', $\mathtt{f}$ for 'feminine', $\mathtt{s}$ for 'singular' and $\mathtt{p}$ for 'plural'. We also have the basic types π_{pn} for the subject when only the person and the number matter and π when person, gender and number do not matter. The basic types o and $\boldsymbol{s}$ stand for 'direct object' respectively for 'sentence in the present'. It is assumed that

$$\pi_{pgn} \to \pi_{pn} \to \pi, \text{ for } p \in \{1,2,3\}, \ g \in \{\mathtt{m},\mathtt{f}\} \text{ and } n \in \{\mathtt{s},\mathtt{p}\} .$$

To analyze a string of words, choose types from the dictionary and concatenate them in the order of the words. The string of words is a sentence of the grammar if and only if the concatenated type has a derivation to the sentence type. For example,

$$\begin{array}{l} \textit{Marie} \quad \textit{examine} \quad \textit{Jean} \\ \text{(MARY EXAMINES JOHN)} \\ \underbrace{(\pi_{3\mathtt{fs}})}\ \underbrace{(\pi^r_{3\mathtt{s}}}\ \boldsymbol{s}\ \underbrace{o^\ell)}\ \underbrace{(o)} \quad \to \boldsymbol{s} \end{array}$$

This derivation is justified by the generalized contractions $\pi_{3\mathtt{fs}}\pi^r_{3\mathtt{s}} \to 1$ and $o^\ell o \to 1$. As customary, the types have been written under the words and the generalized contractions are indicated by under-links.

We illustrate the semantical interpretation by a few examples, following (Preller, 2007b). The sentence

$$\textit{Marie examine Jean} \ (\text{MARY EXAMINES JEAN})$$

is usually rendered in predicate logic by

$$\mathtt{examiner}(\mathtt{marie}, \mathtt{jean}).$$

As usual, transitive verbs like *examiner* are interpreted by binary relations, here embodied by the binary relational symbol $\mathtt{examiner}(x_1, x_2)$. Looking at the type $\pi^r_{3\mathtt{s}} \boldsymbol{s} o^\ell$ of *examine*, we may argue that the basic type $\boldsymbol{s}$ corresponds to the relational symbol and that the non-basic types determine the argument places. We may even go further and make correspond a particular non-basic type to a particular argument place, here ${\pi_{3\mathtt{s}}}^r$ to the first argument place, x_1, and o^ℓ to the second, x_2. Accordingly, the types for proper names, which are just single basic types,

do not introduce argument places and are translated by individual constants. Hence we may add a semantic *translation* for each entry in the dictionary above

$$\begin{array}{lll} \textit{Marie} & : \pi_{3\mathtt{fs}}, o & \mathtt{marie} \\ \textit{Jean} & : \pi_{3\mathtt{ms}}, o & \mathtt{jean} \\ \textit{examine} & : \pi^r_{3\mathtt{s}} \boldsymbol{s} o^\ell & \mathtt{examiner}(x_1, x_2) \end{array} .$$

The translation depends both on the word and its chosen type. For each lexical entry we can create new non-logical symbols or reuse others, introduced earlier.

The reduction of the sentence

$$\begin{array}{llll} \textit{Marie} & \textit{examine} & \textit{Jean} & \\ \underbrace{(\pi_{3\mathtt{fs}})} & \underbrace{(\pi^r_{3\mathtt{s}}\ \boldsymbol{s}\ o^\ell)} & \underbrace{(o)} & \rightarrow \boldsymbol{s} \end{array}$$

suggests that the translating formula $\mathtt{examiner}(\mathtt{marie}, \mathtt{jean})$ can be computed by substitution according to the links. The under-link from $\pi_{3\mathtt{fs}}$ to $\pi^r_{3\mathtt{s}}$ tells us that the constant `marie` translating the basic type $\pi_{3\mathtt{fs}}$ occupies the first argument place x_1 . Similarly, the under-link from o^ℓ to o puts the second constant `jean` into the second argument place.

Generalizing these heuristic considerations, we may agree that a translation of a given lexical entry $word : t_1 \dots t_n$ respects the following rules :

- each basic type t_i is translated by a functional or relational symbol,

- each non-basic type t_i corresponds to an argument-place of at least one functional or relational symbol of the entry,

-the translation of a sentence, via a reduction to the sentence type, is computed by substituting according to the links of the reduction.

Computing the translation of a sentence from the translation of its words makes the translation clearly compositional. Only one rule is needed to explain how the parts are to be composed, namely substitution. We will illustrate this translation mechanism by our sample sentences, beginning with the compound past of transitive words. A warning to the reader : the type assigned below to the past participle of transitive words corresponds to the case when it is used to form the compound past of the active form. In later examples concerning the passive, it will be assigned a different type. No claim to be exhaustive is made of course, new types may always be added without undoing the already recognized sentences due to the conservativity of extensions of pregroup grammars.

Syntax without translation

Suppose we added a new basic type $\boldsymbol{p}$ standing for the past participle and the lexical entries *examiné* : $\boldsymbol{p}\, o^\ell$ and *a* : $\pi^r_{3\mathtt{s}} \boldsymbol{s} \boldsymbol{p}^\ell$ to our dictionary. The augmented dictionary would recognize the sentence *Jean a examiné Marie* (JOHN HAS EXAMINED MARY) by the reduction

$$\begin{array}{llll} \textit{Jean} & \textit{a} & \textit{examiné} & \textit{Marie} \\ \underbrace{(\pi_{3\mathtt{ms}})} & \underbrace{(\pi^r_{3\mathtt{s}} \boldsymbol{s}\, \boldsymbol{p}^\ell)} & \underbrace{(\boldsymbol{p}\, o^\ell)} & \underbrace{(o)} \end{array} .$$

However, the entry *examiné* : $\boldsymbol{p}\, o^\ell$ would correspond to a unary relation. That means that the relation translating a transitive verb in the past would depend only on the object. Moreover, the semantic role of the auxiliary would - correctly - provide the temporal aspect, but depend on the acting individual(s). Therefore, the translation would be a temporal operator that depends on

individuals in opposition to the usual formalizations and interpretations of temporal operators.

Syntax with translation

Adopting the following types 'with translation'

$$\begin{array}{lll} \textit{examiné} : \pi^r \boldsymbol{p} o^\ell & & \texttt{examiner}(x_1, x_2) \\ a \qquad\quad : \pi^r_{3\mathtt{s}} \boldsymbol{s} \boldsymbol{p}^\ell \pi_{3\mathtt{s}} & & \texttt{avoir}(y)\ \texttt{id}(x) \end{array} ,$$

we get the reduction

$$\begin{array}{llllll} \textit{Jean} & a & & \textit{examiné} & \textit{Marie} & \\ (\underbrace{\pi_{3\mathtt{ms}}}) & (\pi^r_{3\mathtt{s}},\ \boldsymbol{s} & \underbrace{\boldsymbol{p}^\ell \underbrace{\pi_{3\mathtt{s}}}) \ (\pi^r \underbrace{\boldsymbol{p}}} & \underbrace{o^\ell)} & \underbrace{(o)} & . \end{array}$$

Now we can correctly interpret the past participle by a binary relation, the same as for other forms of the verb. Next, the type $\pi^r_{3\mathtt{s}} \boldsymbol{s} \boldsymbol{p}^\ell \pi_{3\mathtt{s}}$ for the auxiliary a is now a string of four simple types two of which are basic types, namely $\boldsymbol{s}$ and $\pi_{3\mathtt{s}}$. The other two are the right adjoint $\pi^r_{3\mathtt{s}}$ and the left adjoint $\boldsymbol{p}^\ell$, which correspond to an argument-place x and to an argument place y in this order. Each of the two basic types $\boldsymbol{s}$ and $\pi_{3\mathtt{s}}$ of the entry is associated to a non-logical symbol, namely $\boldsymbol{s}$ to the predicate symbol `avoir` and $\pi_{3\mathtt{s}}$ to the functional symbol `id`. The latter depends on the argument-place x given by $\pi^r_{3\mathtt{s}}$, whereas `avoir` depends on the argument-place y (corresponding to $\boldsymbol{p}^\ell$). This means that the semantic translation of a single word may comprise several logic expressions.

The correspondence between simple types and non-logical symbols is

$$\begin{array}{llllll} a & : & \pi^r_{3\mathtt{s}} & \boldsymbol{s} & \boldsymbol{p}^\ell & \pi_{3\mathtt{s}} \\ & & x & \texttt{avoir} & y & \texttt{id} \\ \textit{examiné} & : & \pi^r & \boldsymbol{p} & o^\ell & \\ & & x_1 & \texttt{examiner} & x_2 & \end{array} .$$

The reduction

$$\begin{array}{llllll} \textit{Jean} & a & & \textit{examiné} & \textit{Marie} & \\ \texttt{john} & x\ \texttt{avoir}\ y\ \texttt{id} & & x_1\ \texttt{examiner}\ x_2 & \texttt{marie} & \\ (\underbrace{\pi_{3\mathtt{ms}}}) & (\pi^r_{3\mathtt{s}},\ \boldsymbol{s} & \underbrace{\boldsymbol{p}^\ell \underbrace{\pi_{3\mathtt{s}}}) \ (\pi^r \underbrace{\boldsymbol{p}}} & \underbrace{o^\ell)} & \underbrace{(o)} & , \end{array}$$

defines the substitutions

$$[x \mid \texttt{jean}],\ [y \mid \texttt{examiner}],\ [x_1 \mid \texttt{id}] \text{ and } [x_2 \mid \texttt{marie}] .$$

The translation of the sentence above is now obtained by substituting according to the underlinks, i.e.

$$\texttt{avoir}(\texttt{examiner}(\texttt{id}(\texttt{jean}), \texttt{marie})).$$

The functional symbol `id` only serves to push the subject from the left side of the auxiliary to the right, i.e. it behaves like the identity function. This is expressed by the non-logical axiom

$$\texttt{id}(x) = x$$

Using the equality $\texttt{id}(\texttt{jean}) = \texttt{jean}$, we derive

$$\texttt{avoir}(\texttt{examiner}(\texttt{jean}, \texttt{marie}))$$

The structure of this expression suggests to read `avoir` as a modal operator applied to an atomic formula. Axioms could be added to the logic to express its temporal meaning, but this goes beyond the scope of our endeavor here. The example of the auxiliary shows that the translation of certain words may comprehend more than one expression of the logic. The predicate symbol `avoir` renders the fact that a sentence is translated by a formula. The functional symbol `id` serves a purpose similar to that of an index in HPSG's : it is used to handle unbounded dependencies.

More generally, the infinitive of a verb will have the same number of arguments as its finite forms, but its type does not depend on the person, gender or number. Hence in the case of the infinitive, we choose the following types and corresponding translations

$$\begin{array}{ll} \textit{avoir} & : \pi^r \boldsymbol{ip}^\ell \pi \quad \texttt{avoir}(y)\ \texttt{id}(x) \\ \textit{examiner} & : \pi^r \boldsymbol{i} o^\ell \quad \texttt{examiner}(x_1, x_2) \end{array} .$$

The logic underlying this semantic interpretation is two-sorted first order logic, one sort for individuals and the other one for sets of individuals, with two primitive relational symbols, namely $\in$ and $=$. It is equivalent to Henkin's system of second order logic with general models, see (van Benthem, 2005).

Before continuing the presentation of the grammar, we want to connect the subject types $\pi^r, \pi^r_{3\mathtt{s}}$ in the past participle or the infinitive with the higher order types of categorial grammars. The usual translation from categorial grammars to pregroup grammars is defined by $A/B \longmapsto AB^\ell$ and $B\backslash A \longmapsto B^r A$. An example why this translation can lead to an overgenerating pregroup grammar is studied in (Moortgat & Oehrle, 2005). It concerns the type of the relative pronoun *that*, for example in the expression *book that Alice found*. The types in Non-associative Lambek calculus with modal operators and their translation into pregroup calculus are

$$\begin{array}{lll} \textit{book} & : n & n \\ \textit{that} & : (n\backslash n)/(\boldsymbol{s}/\Diamond\Box np) & n^r n\, (o*)^{\ell\ell} \boldsymbol{s}^\ell \\ \textit{Alice} & : np & o \\ \textit{found} & : (np\backslash \boldsymbol{s})/np & o^r \boldsymbol{s} o^\ell \end{array} ,$$

where we have written $o*$ for $\Diamond\Box np$ and o for np. The basic types are identical in both grammars. The compound type $\Diamond\Box np$ of the NL-grammar is assimilated to a basic type $o* = \Diamond\Box np$ in the pregroup grammar. Moreover, $o* \to o$ is postulated in (Lambek, 2004). However, this pregroup grammar overgenerates. It recognizes both the grammatical

$$\begin{array}{l} \textit{book} \qquad \textit{that} \quad \textit{Alice}\ \textit{found} \\ \underbrace{(n)}\ \underbrace{(n^r}\ n\ \underbrace{o*^{\ell\ell}\ \underbrace{\boldsymbol{s}^\ell)}\ \underbrace{(o)}\ \underbrace{(o^r}\ \underbrace{\boldsymbol{s}}\ \underbrace{o^\ell)}} \to n\,. \end{array} \tag{1}$$

and, incorrectly, the non-grammatical

$$\begin{array}{l} *\textit{book} \qquad \textit{that} \quad \textit{Alice}\ \textit{found} \quad \textit{it} \quad \textit{and} \\ \underbrace{(\,n)}\ \underbrace{(n^r}\ n\ \underbrace{o*^{\ell\ell}\ \underbrace{\boldsymbol{s}^\ell)}\ \underbrace{(o)}\ \underbrace{(o^r}\ \underbrace{\boldsymbol{s}}\ \underbrace{o^\ell)}\ \underbrace{(o)}\ \underbrace{(o^r}\ \underbrace{o}\ \underbrace{o^\ell)}} \to n\,. \end{array} \tag{2}$$

The derivation responsible for the overgeneration is characterized in *loc.cit.*and used to formulate a rule that excludes this sort of derivations. The result is an enriched grammar with new rules, similar to the constraints on movement in transformational grammars.

A closer look at the difference between the NL-derivation and the pregroup derivation makes it possible to define an ordinary pregroup grammar that recognizes (1), rejects (2) and assigns an appropriate semantic interpretation to (the type of) *that*. In the pregroup derivation (1), the type of *Alice found* is computed directly from the types listed in the dictionary, namely

$$\begin{array}{c} \textit{Alice} \quad \textit{found} \\ \underbrace{(o)} \ \underbrace{(o^r \ \boldsymbol{s} \ o^\ell)} \rightarrow \boldsymbol{s} o^\ell \end{array} \tag{3}$$

The modal operators, on the contrary, transform $(Alice \circ found) \vdash \boldsymbol{s}/np$ to

$$(Alice \circ found) \vdash \boldsymbol{s}/\Diamond\Box np \tag{4}$$

before it is concatenated with the type of *that* :

$$\dfrac{\dfrac{book}{n} \qquad \dfrac{that}{(n\backslash n)/(\boldsymbol{s}/\Diamond\Box np)} \qquad \dfrac{\dfrac{\dfrac{\dfrac{Alice}{np} \qquad \dfrac{found}{(np\backslash\boldsymbol{s})/np} \qquad \dfrac{\Box np \vdash \Box np}{\Diamond\Box np \vdash np}(\backslash E)}{found \circ \Diamond\Box np \vdash np\backslash\boldsymbol{s}}}{Alice \circ (found \circ \Diamond\Box np) \vdash \boldsymbol{s}}}{(Alice \circ found) \circ \Diamond\Box np \vdash \boldsymbol{s}}}{Alice \circ found \vdash (\boldsymbol{s}/\Diamond\Box np)}}{that \circ (Alice \circ found) \vdash (n\backslash n)}}{book \circ (that \circ (Alice \circ found)) \vdash n} \tag{5}$$

As pregroup grammars only use types from the dictionary, we must anticipate the type (4), by adding $found : o^r \boldsymbol{s} o*^\ell$ to our dictionary, but take care to keep the two basic types o and $o*$ unrelated. Hence $o \not\rightarrow o*$ in the revised pregroup grammar. The entry $found : o^r \boldsymbol{s} o^\ell$ is retained, so that the pregroup dictionary now lists

$$found : o^r \boldsymbol{s} o^\ell, o^r \boldsymbol{s} o*^\ell \ .$$

As now $o*$ is isolated in the set of basic types, we may rename $o*^{\ell\ell}$ as $\bar{o}$ and therefore change $o*^\ell$ to $\bar{o}^r$ without changing derivations. The resulting dictionary

$$\begin{array}{lll} book & : n & n \\ that & : (n\backslash n)/(\boldsymbol{s}/\Diamond\Box np) & n^r n \, \bar{o} \boldsymbol{s}^\ell \\ Alice & : np & o \\ found & : (np\backslash\boldsymbol{s})/np & o^r \boldsymbol{s} o^\ell, \ o^r \boldsymbol{s} \bar{o}^r \end{array}$$

is strongly equivalent to the dictionary before the replacement.[1] More generally it can be shown that due to compactness an arbitrary pregroup dictionary is strongly equivalent to one with no iterated adjoints. The latter has a semantic interpretation, for example the 'dummy' $\bar{o}$ in the type of *that* will play the role of a temporary name for the set of entities satisfying the following relative clause. If the auxiliary *avoir* is seen as a map from predicates to predicates, a 'dummy' is introduced in its pregroup type by the translation as indicated above.

[1]Following the categorial type, we have used here the same symbol for noun phrases, regardless whether they occur in subject or object position. In the rest of the paper, we continue to split the type of a noun phrase into π and o as indicated earlier. Both are interpreted as (subsets of) individuals.

2 Distant agreement in French

French clitics have already been studied with pregroup grammars in (Bargelli-Lambek), but without agreement. Our analysis differs from that given in the latter for two reasons. First of all, we want to avoid the meta-rule used there and base the analysis inside an ordinary pregroup grammar. The other reason is that we prefer to think of clitics as designating individuals or sets of individuals, not operators on relations. Due to the restricted space, only agreement with the preverbal personal pronoun in the role of a direct object is presented in some detail.

In the compound past of the active form, the past participle agrees in gender and number with the direct object clitic. If the verb is in passive mode or forms its compound past with the auxiliary *être*, the past participle agrees in gender and number with the subject. Reflexive pronouns, which are perverbal in French, agree with the subject.

We add to the basic types of the preceding section new basic types for direct object clitics o_{pgn}, depending on the features of person $p = 1, 2, 3$, gender $g = \mathtt{m}, \mathtt{f}$ and number $n = \mathtt{s}, \mathtt{p}$. The 'dummies' $\hat{o}_{pgn}$ and $\hat{o}$ will capture distant dependencies. The types $\hat{o}_{gn}$ are used if only gender and number matter, but not the person. The dependence of the type of the clitic on the person makes it possible to avoid non grammatical combinations of two clitics, like **me lui*, but we do not pursue this topic here. We assume

$\hat{o}_{pgn} \to \hat{o}_{gn} \to \hat{o}$, $o_{pgn} \to \hat{o}_{gn} \to \hat{o}$, for $p = 1, 2, 3$, $g = \mathtt{m}, \mathtt{f}$ and $n = \mathtt{s}, \mathtt{p}$.

We use the following sentences to illustrate how pregroup grammars can handle syntactical and semantical agreement :

Marie les examine	*Marie s'examine*
(MARY EXAMINES THEM)	(MARY EXAMINES HERSELF)
Marie les a examinés	*Marie s'est examinée*
(MARY HAS EXAMINED THEM)	(MARY HAS EXAMINED HERSELF)
Marie doit les examiner	*Marie doit s'examiner*
(MARY MUST EXAMINE THEM)	(MARY MUST EXAMINE HERSELF)
Marie doit les avoir examinés	*Marie doit s'être examinée*
(MARY MUST HAVE EXAMINED THEM)	(MARY MUST HAVE EXAMINED HERSELF)
Marie est examinée par Jean	*Marie est examinée*
(MARY IS EXAMINED BY JOHN)	(MARY IS EXAMINED)

The lexical entries for the personal pronoun *les* and the reflexive pronoun *s'* are

les :	$o_{3g\mathtt{p}}$	$\mathtt{Les}$
s' :	$\pi^r_{3gn}\pi_{3gn}\hat{o}_{3gn}$	$\mathtt{id}(x)\ \mathtt{id}(x)$

, where $g \in \{\mathtt{m}, \mathtt{f}\}$, $n \in \{\mathtt{s}, \mathtt{p}\}$.

If the context permits, the set of values for the subscripts p, g, n is omitted.

Note that the same subscripts g and n appear both in the right adjoint π^r_{3gn} and in the dummy type $\hat{o}_{3gn}$. Hence, the values of the features of gender and number of the subject are identical to those of the dummy object. The anaphoric content is captured by the occurrence of two basic types, π_{3gn} and $\hat{o}_{3gn}$, which both are translated by the unary functional symbol $\mathtt{id}$. The argument-place x corresponds to π^r_{3gn}. The effect is to repeat the entity which will be substituted for x, because $\mathtt{id}(x) = x$ holds in the logic.

In simple tenses, clitics do not require agreement with the following verb. Their preverbal position makes it necessary to assign a new type to the verb. These new entries are added to the ones given in the preceding section :

$$\begin{array}{lll} \textit{examiner}: & \hat{o}^r\pi^r\boldsymbol{i} & \texttt{examiner}(z_2, z_1) \\ \textit{examine} \ : & \hat{o}^r\pi^r_{3\texttt{s}}\boldsymbol{s} & \texttt{examiner}(z_2, z_1) \end{array} .$$

Note that the order of the variables and the non-logical symbols in the translation is not arbitrary. It is used to code the correspondence of these symbols with the simple types of the entry. In the new entries, the first variable z_1 corresponds to $\hat{o}^r$ and the second argument place z_2 to π^r respectively $\pi^r_{3\texttt{s}}$. A simple convention will make the explicit definition of the correspondence between simple types in the entry and non-logical symbols in the translation superfluous. It suffices to count the non-basic types from left right and assign them a new variable in each occurrence. Similarly, the non-logical symbols correspond to the basic types in their order of occurrence.

Then we find the following reductions

Marie les examine	*Marie s' examine*
(MARY EXAMINES THEM)	(MARY EXAMINES HERSELF)
$(\pi_{3\texttt{fs}})\ \underbrace{(o_{3gp})\ (\hat{o}^r\pi^r_{3\texttt{s}}}\ \boldsymbol{s})\,, \quad g = \texttt{m}, \texttt{f}$	$(\pi_{3\texttt{fs}})\ \underbrace{(\pi^r_{3\texttt{fs}},\ \pi_{3\texttt{fs}}\hat{o}_{3\texttt{fs}})\ (\hat{o}^r\pi^r_{3\texttt{s}}}\ \boldsymbol{s})\,.$

As g can take two values, the left hand display corresponds to two different type assignments, differing by $o_{3\texttt{mp}}$ and $o_{3\texttt{fp}}$ for the clitic *les*. The reduction itself remains unchanged, the set of links is the same for both type assignments.

Note that the semantic difference between the left and right hand sentences above is correctly captured by the reductions. The left hand reductions define the translation

$$\texttt{examiner}(\texttt{marie}, \texttt{Les})\,,$$

whereas the right hand reduction gives

$$\texttt{examiner}(\texttt{id}(\texttt{marie}), \texttt{id}(\texttt{marie}))\,.$$

As $\texttt{id}(x) = x$, the latter translation is equivalent to

$$\texttt{examiner}(\texttt{marie}, \texttt{marie})\,.$$

The type of the reflexive pronoun depends on the person to avoid non-sentences like **Tu s'examine*(YOU EXAMINE HIMSELF). Indeed, $tu : \pi_{2gs},\ g = \texttt{m}, \texttt{f}$ and $\pi_{2gs}\pi^r_{3gs} \not\to 1$.

In the compound past, the clitic is separated from its verb by the auxiliary. The auxiliary does not show the relevant features by its form, but it passes them to the following word(s). The lexical entries below model this behavior by 'remembering' types.

$$\begin{array}{lllll} \textit{examiné} & : & \hat{o}^r_{\texttt{ms}}\pi^r\boldsymbol{p} & \texttt{examiner}(x_2, x_1) & \\ \textit{examinée} & : & \hat{o}^r_{\texttt{fs}}\pi^r\boldsymbol{p} & \texttt{examiner}(x_2, x_1) & \\ \textit{examinés} & : & \hat{o}^r_{\texttt{mp}}\pi^r\boldsymbol{p} & \texttt{examiner}(x_2, x_1) & \\ \textit{examinées} & : & \hat{o}^r_{\texttt{fp}}\pi^r\boldsymbol{p} & \texttt{examiner}(x_2, x_1) & \\ \textit{avoir} & : & o_{pgn}{}^r\pi^r\boldsymbol{ip}^\ell\pi\hat{o}_{gn} & \texttt{avoir}(y_3) & \texttt{id}(y_2) \quad \texttt{id}(y_1) \\ \textit{a} & : & o_{3gn}{}^r\pi^r_{3\texttt{s}}\boldsymbol{sp}^\ell\pi\hat{o}_{gn} & \texttt{avoir}(y_3) & \texttt{id}(y_2) \quad \texttt{id}(y_1) \\ \textit{être} & : & \hat{o}^r_{pgn}\pi^r\boldsymbol{ip}^\ell\pi\hat{o}_{gn} & \texttt{être}(y_3) & \texttt{id}(y_2) \quad \texttt{id}(y_1) \\ \textit{est} & : & \hat{o}^r_{3gs}\pi^r_{3\texttt{s}}\boldsymbol{sp}^\ell\pi\hat{o}_{\texttt{gs}} & \texttt{être}(y_3) & \texttt{id}(y_2) \quad \texttt{id}(y_1) \end{array} .$$

Here too, we have followed our convention that the variables respectively non-logical symbols correspond to the non-basic types respectively basic types in their order of occurrence. For

example, consider the last four entries above. The variables y_1 and y_2 correspond to the right adjoints o^r and π^r, with the appropriate subscripts, with or without hat. The left adjoint $\boldsymbol{p}^\ell$ corresponds to the variable y_3. The basic type $\boldsymbol{i}$ respectively $\boldsymbol{s}$ is translated by the relational symbol, the basic types π and $\hat{o}_{gn}$ are translated by the functional symbol $\mathtt{id}$. Choosing the value $g = \mathtt{f}$ in the type $o_{3\mathtt{fp}}$ for *les*, the sentence *Marie les a examinées* is recognized by the following reduction

$$\begin{array}{llll} \textit{Marie} & \textit{les} & \textit{a} & \textit{examinées} \\ \underbrace{(\pi_{3\mathtt{fs}})}\ \underbrace{(o_{3\mathtt{fp}})\ (o^r_{3\mathtt{fp}}\pi^r_{3\mathtt{s}})}\ \boldsymbol{s}\ \underbrace{\boldsymbol{p}^\ell\ \pi\ \underbrace{\hat{o}_{\mathtt{fp}})\ (\hat{o}^r_{\mathtt{fp}}}\pi^r\ \boldsymbol{p})}} & & & \end{array} .$$

If the clitic is a reflexive pronoun, the auxiliary in the compound tense is *être*. The past participle agrees in gender and number with the clitic if the latter is the direct object. Hence the type of *être* is similar to that of *avoir*, except that it is tailored to the reflexive pronoun, and therefore uses the dummy object types.

$$\begin{array}{llll} \textit{Marie} & \textit{s'} & \textit{est} & \textit{examinée} \\ \underbrace{(\pi_{3\mathtt{fs}})\ (\pi^r_{3\mathtt{fs}}}\ \underbrace{\pi_{3\mathtt{fs}}\ \hat{o}_{3\mathtt{fs}})\ (\hat{o}^r_{3\mathtt{fs}}\pi^r_{3\mathtt{s}})}\ \boldsymbol{s}\ \underbrace{\boldsymbol{p}^\ell\ \pi\ \underbrace{\hat{o}_{\mathtt{fs}})\ (\hat{o}^r_{\mathtt{fs}}}\pi^r\ \boldsymbol{p})} & & & \end{array} .$$

Note that the hat on the direct object in the type of *est* prevents **Marie l'est examinée* and **Marie s'a examiné* as $o_{3\mathtt{fs}} \not\to \hat{o}_{3\mathtt{fs}}$ and $\hat{o}_{3\mathtt{fs}} \not\to o_{3\mathtt{fs}}$.

The translation of the latter sentence is

$$\mathtt{\hat{e}tre}(\mathtt{examiner}(\mathtt{marie}, \mathtt{marie}))\ .$$

Whereas the auxiliaries *avoir* and *être* 'remember' the features of the object, the modal verbs 'remember' the features of the subject. The clitic is positioned between the modal verb and the verb of which it is the complement.

$$\begin{array}{lll} \textit{devoir}: & \pi^r_{pgn}\boldsymbol{i}\boldsymbol{i}^\ell\pi_{pgn} & \mathtt{devoir}(y)\ \mathtt{id}(x) \\ \textit{doit} & : \pi^r_{3g\mathtt{s}}\boldsymbol{s}\boldsymbol{i}^\ell\pi_{3g\mathtt{s}} & \mathtt{devoir}(y)\ \mathtt{id}(x) \end{array}$$

where $p = 1, 2, 3;\ g = \mathtt{m}, \mathtt{f};\ n = \mathtt{s}, \mathtt{p}$. In these entries, the unary relational symbol $\mathtt{devoir}$ translates the basic types $\boldsymbol{i}$ and $\boldsymbol{s}$. The unary functional symbol $\mathtt{id}$ translates the basic type π_{pgn}. The variable y corresponds to $\boldsymbol{i}^\ell$ and x to π^r_{pgn}. The reason why the type of the modal verbs depends on the gender becomes evident when they are used in combination with the compound past. For example

$$\begin{array}{llllll} \textit{Marie} & \textit{doit} & \textit{s'} & \textit{être} & \textit{examinée} & \\ (\pi_{3\mathtt{fs}})(\pi^r_{3\mathtt{fs}}\ \boldsymbol{s}\ \underbrace{\boldsymbol{i}^\ell\ \pi_{3\mathtt{fs}})(\pi^r_{3\mathtt{fs}}}\ \underbrace{\pi_{3\mathtt{fs}}\ \hat{o}_{3\mathtt{fs}})(\hat{o}^r_{3\mathtt{fs}}\ \pi^r}\ \boldsymbol{i}\ \underbrace{\boldsymbol{p}^\ell\ \pi\ \hat{o}_{\mathtt{fs}})(\hat{o}^r_{\mathtt{fs}}\ \pi^r\ \boldsymbol{p}}) & & & & & \end{array} .$$

The reader may verify that the translation renders the correct meaning and check that the non-sentences **Marie doit s'être examiné*, **Marie doit s'être examinés* and **Marie doit s'être examinées* have no reduction to the sentence type.

3 Conclusion

Pregroup dictionaries use more basic types and entries per word than categorial grammars. This is the price to pay for reducing computation to a single rule, namely generalized contraction.

In spite of this 'explosion' of types, dictionaries like the one presented here have parsing algorithms which are linear when given strings of lexical entries, see (Preller, 2007a). Exploiting certain regularities of features rendering 'lazy type assignment' possible, this result is improved in forthcoming work by (Preller & Prince, 2006) : there is a linear algorithm which for a given string of words finds a parsing, i . e. a reduction to the sentence type, if and only if the string of words is a sentence.

Remerciements

The author would like to thank Violaine Prince for her insistence to tackle anaphora with pregroup grammars.

Références

LAMBEK J. (1999). Type grammar revisited. In A. L. ET AL., Ed., *Logical Aspects of Computational Linguistics*, volume 1582 of *LNAI*, p. 1–27. Springer.

LAMBEK J. (2004). A computational algebraic approach to english grammar. *Syntax*, **7**(2), 128–147.

MOORTGAT M. & OEHRLE R. T. (2005). Pregroups and type-logical grammar : Searching for convergence. In C. CASADIO, P. SCOTT & R. SEELEY, Eds., *Language and Grammar—Studies in Mathematical Linguistics and Natural Language*. CSLI Publications.

PRELLER A. (2007a). Linear processing with pregroup grammars. *Studia Logica*, **forthcoming**.

PRELLER A. (2007b). Toward discourse representation via pregroup grammars. *Language Logic and Information*, **16**(2), 173–194. published on-line 02.02.2007, DOI 10.1007/s10849-0006-9033-y.

PRELLER A. & PRINCE V. (2006). *Pregroup grammars with linear parsing : long distance dependency of clitics in French*. Rapport interne, LIRMM, Universtité de Montpellier.

VAN BENTHEM J. (2005). Guards, bounds and generalized semantics. *Journal of Language, Logic and Information*, **14**, 263–279.

Atelier TALN-2007

5 au 8 juin 2007, Toulouse, France

Actes de l'atelier

RECONSTRUIRE LA LANGUE

DANS LES COMMUNICATIONS ALTERNATIVES

ET AUGMENTÉES

Éditeur scientifique

Maryvonne ABRAHAM

Organisation de la conférence

ENST Bretagne & IRIT (UMR 5505)

Comité d'organisation

Maryvonne ABRAHAM*	*GET - ENST-Bretagne*
Jean-Yves ANTOINE	*Dir. IUP BLois-Tours Université*
Philippe BLACHE	*LPL - Université de Provence*
Philippe BOISSIÈRE	*IRIT - Toulouse*
Denis MAUREL	*LI Université François Rabelais - Tours*
Nadine VIGOUROUX	*IRIT - Toulouse*

Comité de programme

Maryvonne ABRAHAM*	*GET - ENST-Bretagne*
Philippe BOISSIÈRE	*IRIT - Toulouse*
Nadine VIGOUROUX	*IRIT - Toulouse*

* Présidente

Session
Communications orales

Le module de reformulation iconique de la Plateforme de Communication Alternative

Philippe BLACHE, Stéphane RAUZY
Laboratoire Parole et Langage
CNRS & Université de Provence

{pb,tephane.rauzy}@lpl.univ-aix.fr

Résumé. Nous présentons dans cette contribution le système de reformulation iconique implanté dans le logiciel d'aide à la communication pour personnes handicapées Plateforme de Communication Alternative (PCA). Il s'agit de générer, à partir d'un message composé d'une séquence d'icônes, une phrase en langage naturel syntaxiquement et sémantiquement correcte. Le module de reformulation de la PCA répond à une double contrainte. Le système doit d'une part proposer une interprétation couvrante, en terme du nombre et du type de messages effectivement reformulés. D'autre part, l'utilisateur généralement non-expert en linguistique doit pouvoir enrichir son matériel lexical par ajout de nouveaux items. Le lexique doit ainsi porter des informations linguistiques minimales accessibles à tous (l'utilisateur, sa famille ou le personnel accompagnant) via une interface simplifiée. Cette double contrainte conditionne en pratique le choix des règles de reformulation implémentées dans le système et les performances du processus de reformulation.

Abstract. We present the reformulation system embedded in "Plateforme de Communication Aternative", an assistive communication software destinated to impaired persons. The objective is to transform an input message composed of a sequence of iconic items in a well-formed output sentence, both with regard to syntax and semantics. The constrain is herein two-fold. On one hand, the system has to propose a maximal coverage interpretation in the space of reformulated entries. On the other hand, the end user generally non-expert in linguistics may wish to enrich its lexical material by adding new iconic items. The lexicon must then bring minimal linguistic informations accessible to everyone (the end user, its familly or support staff) through a simplified interface. This double constrain governs in practice the choice of the reformulation rules implemented in the system and determines indeed the performances of the reformulation process.

Mots-clefs : communication assistée pour personnes handicapées,communication non verbale,système de reformulation.

Keywords: assistive communication for impaired persons,non verbal communication, reformulation system.

1 Introduction

La communication alternative désigne un ensemble d'outils d'aide à la communication pour des personnes handicapées atteintes dans leur motricité et leur capacité de production de parole. Il s'agit par exemple de patients atteints de pathologies neuro-dégénératives totalement paralysantes ou encore de personnes victimes d'accidents vasculaires cérébraux. Ces patients ne gardent le contrôle que de quelques muscles (comme la paupière) et ne peuvent plus parler. Pour d'autres pathologies, certains types d'aphasies par exemple, les capacités linguistiques et cognitives sont affectées et des stratégies alternatives comme la communication non verbale à base d'icônes doivent être utilisées. L'objectif de ce type de système est de permettre à l'utilisateur d'améliorer voire rétablir la possibilité de communication avec son entourage en offrant la possibilité de composer des messages, de piloter un système de synthèse de parole ou encore de désigner des objets ou des actions. Il s'agit donc de prendre en compte les besoins effectifs de ces utilisateurs dans une situation réelle de communication, et d'intégrer des modalités multiples d'interaction pour le support de la communication et le contrôle de l'environnement (voir par exemple Vaillant (1997) et Brangier&Gronier (2000)).

L'aide à la communication de personnes handicapées est un problème majeur, mais qui peut aujourd'hui bénéficier de la maturité technologique des travaux menés dans le domaine de la linguistique, la linguistique-informatique et la psychologie cognitive. Les réponses apportées à ce jour ne sont pas totalement satisfaisantes, notamment pour ce qui concerne les modalités d'interaction entre l'utilisateur handicapé et son environnement humain ou électronique (voir par exemple Maurel et al. (2000)).

Quelques systèmes d'aide à la communication proposent une solution globale qui intègre un module de communication verbale et un module de communication non verbale. Citons par exemple pour le verbal : WiViK, clavier virtuel avec prédiction et défilement en option, permettant également le contrôle du système ; Eurovocs Suite, claviers virtuels et prédiction de mots basée sur un dictionnaire contenant 35.000 formes. Et pour la communication non verbale : Clicker 4, outil d'aide à la communication à base d'icônes ; Mind Express, un système de communication non verbale à base d'icônes qui intègre une reformulation rudimentaire. Enfin Axelia, certainement le logiciel le plus avancé pour le français, qui traduit une suite d'icônes en une phrase de la langue naturelle, accessible via une interface graphique évoluée. Axelia base sa reformulation sur l'application du modèle de la grammaire applicative et cognitive (voir à ce sujet Abraham (2000), Abraham (2006)). Il est destiné aux jeunes enfants I.M.C. (Infirme Moteur Cérébraux) et aphasiques.

Il existe enfin un certain nombre d'applications expérimentales développées dans le milieu académique : par exemple, VITIPI (Boissière et al. (2000)), HandiAS (Le Pevedic (1997)) ou Kombe (Pasero&Sabatier (1995)), mais qui ne sont pas véritablement distribués au grand public.

Le système développé au Laboratoire Parole et Langage et distribué depuis début 2004, la Plateforme de Communication Alternative (PCA), intègre un certain nombre de caractéristiques d'homogénéité et de généricité nécessaires à toute bonne communication assistée (voir Copestake (1997), Blache&Rauzy (2003), Bellengier et al. (2004), Blache&Rauzy (2004)). Le logiciel PCA permet la composition assistée de messages selon deux modes principaux : le mode verbal et le mode non-verbal. Ces deux types de composition sont accessibles par le clavier, la souris, ou une procédure de défilement, selon le degré de motricité des utilisateurs.

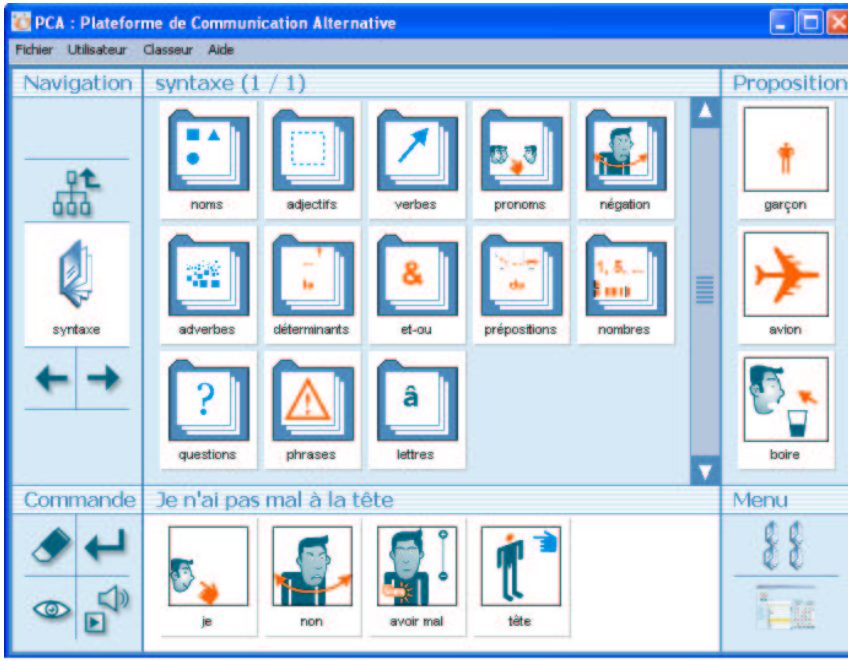


FIG. 1 – L'interface de la Plateforme de Communication Alternative en mode non verbal

La composition en mode verbal s'effectue à l'aide d'un clavier orthographique statique complété par un clavier dynamique de proposition de mots. Le moteur de prédiction implanté dans PCA utilise un lexique très couvrant du français (320 000 formes fléchies, voir VanRullen et al. (2005)) et propose une prédiction contextuelle incluant l'information sur les traits morphosyntaxiques associées aux entrées du lexique ainsi qu'un modèle utilisateur qui prend en compte les habitudes langagières de l'utilisateur par apprentissage.

La composition en mode non verbal s'effectue à l'aide d'un clavier d'icônes (voir figure 1). La base d'icônes générale partagée par tous les utilisateurs regroupe environ 750 pictogrammes qui ont été dessinés à partir d'une chartre graphique et sémantique élaborée par le Laboratoire Parole et Langage, et testée par de nombreux utilisateurs. Elle couvre des besoins communicationnels variés. La base comprend environ 200 verbes (les verbes les plus courants et des verbes spécialisés utilisés par exemple dans le domaine médical), environ 200 noms communs (désignant des objets, des lieux, des personnes, etc.), une cinquantaine d'adjectifs, les pronoms, les adverbes, les déterminants et les prépositions les plus courants, et les nombres. La base comprend de plus les icônes représentant les lettres et les phonèmes qui permettent de créer des claviers alphabétiques ou phonétiques. Chaque utilisateur pourra ensuite créer et ajouter, via une interface facile d'accès, ses propres icônes (à partir de photos numériques par exemple).

Nous décrivons dans la section suivante le système de reformulation iconique implanté dans la version non verbale du logiciel PCA, c'est-à-dire le module qui génére, à partir d'une séquence d'icônes, une phrase en langage naturel syntaxiquement et sémantiquement correcte.

2 Le module de reformulation iconique

La question de la reformulation en langage naturel d'un message composé d'une séquence d'icônes a été adressée par de nombreux auteurs (voir par exemple McCoy (1997), Abraham (2000), Abraham (2006), Bellengier et al. (2006)). Deux problèmes doivent être abordés. D'une part, quelles sont les informations syntaxiques et sémantiques à associer à chaque icône, ou autrement dit, quelles informations nécessaires à la génération doit-on faire porter sur le lexique ? D'autre part, comment gérer simultanément les contraintes syntaxiques et les contraintes sémantiques dans le cas d'une entrée qui est de fait incomplète, certaines informations étant absentes de la séquence à traiter ?

2.1 Les informations linguistiques nécessaires à la génération

La finesse des informations linguistiques associées aux items du langage conditionne en pratique les performances du module de reformulation. Dans le cas d'un système fermé par exemple, c'est-à-dire un système ne permettant pas l'ajout de nouveau matériel lexical, une description aussi fine que possible des propriétés syntaxiques et sémantiques de chaque item est souhaitable. Dans le cadre de la communication non verbale proposée par la PCA, nous avons opté pour un autre choix. Il nous est apparu indispensable que l'utilisateur puisse enrichir son matériel lexical par des additions ou des modifications de la base d'icônes fournie par défaut. L'ouverture du système conditionne sévèrement les choix d'implémentation des règles de reformulation. En effet, l'utilisateur ou la personne l'accompagnant auront à renseigner, pour chaque nouvel item créé, les informations nécessaires au module de reformulation en langage naturel. La formulation des informations linguistiques doit ainsi être assez simple pour être accessible à tous (c'est-à-dire à des personnes ne possédant pas une expertise particulière dans le domaine de la linguistique).

Nom	Les noms propres et les noms communs
Verbe	Les verbes
Locution verbale	Les verbes suivis d'une locution, ex. "prendre garde"
Adjectif	Les adjectifs qualificatifs
Préposition	Les prépositions, ex. "à", "de", "avec", etc.
Adverbe	Les adverbes, ex. "souvent", "très", "facilement"
Déterminant	Les articles définis, indéfinis, démonstratifs, possessifs
Pronom	Les pronoms personnels et démonstratifs
Et/ou	Les conjonctions de coordination "et" et "ou"
Négation	La négation "ne...pas"
Question	Les unités interrogatives du type "Quand", "Comment", "Qui", etc.
Nombre	Les chiffres et les nombres
Lettre	Les lettres ou groupes de lettres
Phrase	Les phrases ou parties de phrases qui ne sont pas reformulées

FIG. 2 – La liste des catégories syntaxiques du module de reformulation de PCA

Le module de reformulation implanté dans PCA se base ainsi sur un lexique comportant des informations linguistiques minimales. A chaque icône est associée une des catégories syntaxiques listées figure 2. Les informations sémantiques associées à chaque entrée du lexique sont très limitées. Il s'agit principalement de spécifier la nature des noms communs (personne, objet, lieu

ou transport) et les prépositions associées aux verbes. Deux exemples de l'interface de saisie des informations de reformulation sont présentés figure 3.

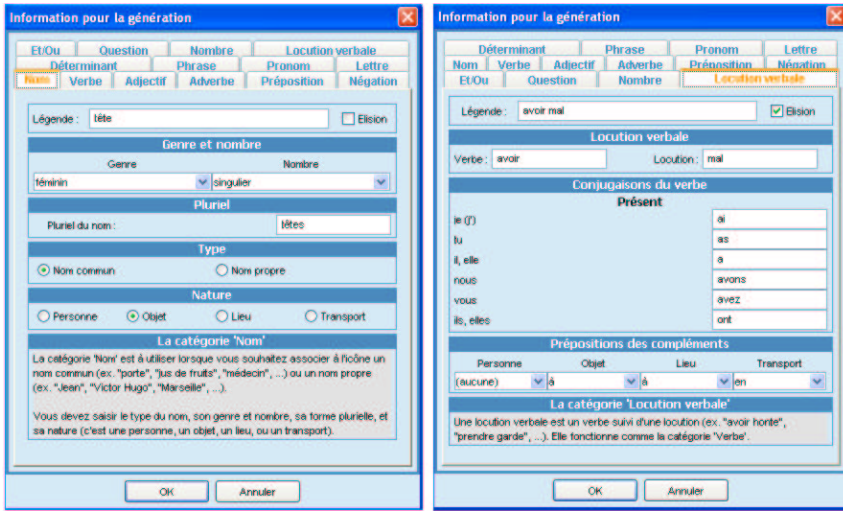


FIG. 3 – L'interface de saisie des informations nécessaires à la reformulation. Lorsque un nouvel item est ajouté au matériel lexical de l'utilisateur, les champs correspondant à la catégorie syntaxique de l'item doivent être renseignés.

2.2 Le processus de génération

La chaîne de traitement entrant dans l'algorithme de génération comporte 4 étapes :

- Etape 1 : Chaque icône de la séquence est traduite, via le lexique, en une catégorie syntactico-sémantique et son label associé.
- Etape 2 : Les catégories sont regroupées en pré-syntagmes (le pré-syntagme est une version sous-déterminée du syntagme final, i.e correctement et totalement reformulé).
- Etape 3 : Les pré-syntagmes sont traités comme des arbres adjoints sur-spécifiés (voir section suivante), ce mécanisme permettant d'instancier complétement les traits syntaxiques des catégories terminales.
- Etape 4 : Les règles de gestion des phénomènes linguistiques particuliers sont appliquées (élision, contractions, déplacements, etc.).

2.3 Les arbres adjoints sur-spécifiés

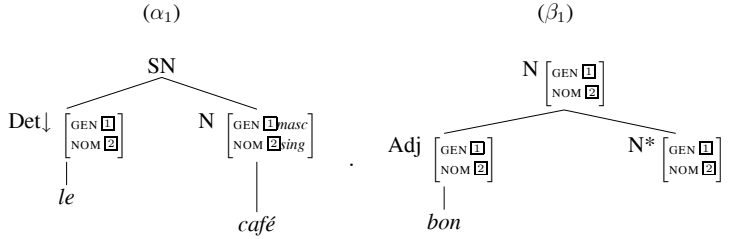
Dans l'étape 3 du proccessus de génération, des informations morpho-syntaxiques de plus haut niveau sont associés aux icônes afin de prendre en compte le contexte de production. Il est également nécessaire de proposer des mécanismes de gestion de ces informations. La solution décrite ici repose sur une adaptation des grammaires d'arbres adjoints (TAG, cf. (Joshi87)). Dans notre approche, les labels sont les ancres d'arbres élémentaires dans lesquels nous ajoutons

un mécanisme de valeurs par défaut permettant d'une part l'insertion de mots manquants et d'autre part la gestion de la flexion.

2.3.1 Les arbres élémentaires associés aux icônes

Les arbres élémentaires utilisés sont, classiquement, de deux types (arbres initiaux et arbre auxiliaires) en fonction du rôle qu'il jouent dans la construction de la structure. Les têtes de ces arbres, correspondant aux étiquettes des icônes, sont des lemmes qui devront, nous le verront plus loin, être fléchis. Leur structure de traits peut être indexée, ce qui autorise la représentation du partage de structure.

Une première caractéristique essentielle distingue cependant nos arbres élémentaires : ils peuvent contenir plusieurs feuilles lexicales, en plus de la tête. Cette propriété nous permettra d'indiquer par exemple les spécifieurs par défaut.



L'arbre initial (α_1) indique un noeud à substitution pour lequel un descendant est spécifié. Il s'agit d'indiquer de cette façon la valeur des compléments éventuellement non spécifiés dans la suite d'icônes. En l'occurrence, cette valeur permettra de rétablir un déterminant s'il est absent. La feuille constitue donc une valeur par défaut, elle ne correspond pas à une forme mais, nous le verrons dans la section suivante, à un ensemble de formes possibles. La gestion de l'accord est assurée par un mécanisme de partage de valeur entre les traits concernés (indiqué par la coindexation).

L'arbre auxiliaire quant à lui permet de préciser les adjoints potentiels. Dans cet arbre, il faut remarquer la gestion simplifiée de l'accord qui consiste seulement à reporter du noeud racine au modifieur les traits d'accord et à les copier sur le noeud pied. On notera au passage que les traits d'accord de l'adjectif ne sont pas spécifiés : comme dans le cas du déterminant de l'arbre (α_1), l'ancre lexicale correspond au lemme, non à la forme. Il est donc nécessaire de préciser le fonctionnement de la flexion permettant de calculer la forme générée.

2.3.2 Les valeurs de traits faibles et fortes

Le mécanisme de base consiste à spécifier d'une part des lemmes par défaut, mais également des formes par défaut pour ces lemmes ainsi que des valeurs de traits par défaut. Cependant, un mécanisme supplémentaire est nécessaire pour gérer la substitution de traits permettant la propagation de l'accord ou le rétablissement de formes correctement accordées. Les exemples suivants illustrent ce phénomène. Ils indiquent à gauche de la flèche les étiquettes des icônes utilisées (avec leur valeur lexicale par défaut) et à droite la reformulation.

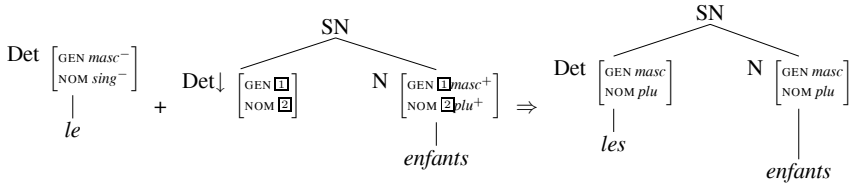
(1) a. le + enfants → les enfants
b. la + homme → l'homme
c. les + homme → les hommes

Dans ces exemples les étiquettes des icônes ne correspondent pas à des formes correctement accordées. Il est donc nécessaire d'établir un mécanisme qui, en plus des valeurs par défaut, permettra à certains traits de se propager.

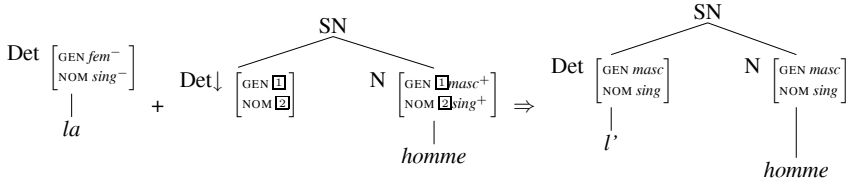
Nous introduisons pour cela deux types de valeurs : *forte* et *faible*. Les valeurs fortes correspondent à des valeurs généralement marquées (en particulier du point de vue morphologique) et qui devront se propager : les valeurs faibles ont vocation à éventuellement être substituées par des valeurs fortes.

Nous notons par la suite les valeurs faibles par un exposant négatif ([GEN masc$^-$]) et les fortes avec un exposant positif. En cas d'unification entre deux traits, l'un comportant une valeur faible et l'autre une valeur forte, cette dernière remplacera la faible à l'issue de la substitution. Ce mécanisme permet de traiter directement les valeurs marquées.

Dans l'exemple (1a), la substitution entraîne un conflit de valeur du trait de nombre. La valeur forte du trait de nombre de l'arbre *enfants* se propage au déterminant qui comporte une valeur faible de trait nombre. Le résultat est donc la propagation du pluriel au déterminant. Les arbres suivants illustrent le mécanisme (nous ne notons pas ici, pour des raisons de lisibilité, les valeurs par défaut associées aux substantifs, qui n'interviennent pas dans le traitement des ces exemples).



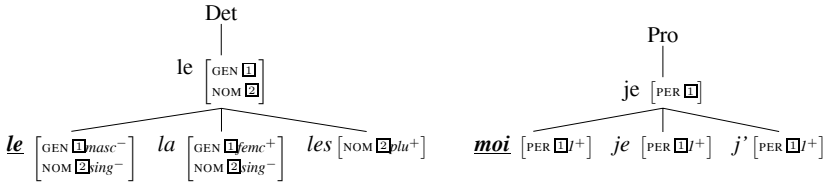
Les arbres suivants illustrent un mécanisme comparable concernant le trait de genre : le trait genre du nom étant fort, il se propagera vers le déterminant. Celui-ci étant faible, la valeur du trait genre du déterminant deviendra masculin, la forme correspondante sera donc forte.



2.3.3 La flexion

Le mécanisme de propagation de traits forts entraînent, nous l'avons vu dans les exemples précédents, une modification éventuelle de la forme du lemme de l'étiquette de l'icône. Ce pro-

cessus est pris en charge par un mécanisme particulier : chaque lemme est associé à un certain nombre de formes, l'une d'entre elle étant une forme par défaut. Lors de la sélection d'une icône, la valeur par défaut (et ses traits associés) est affichée. Nous introduisons une nouvelle relation de sélection entre le lemme et ses formes, distincte d'une relation de dominance.



Le principe consiste à établir une relation fluctuante entre le lemme et l'une des ses formes. La forme affichée sera celle dont les traits s'unifient avec ceux du lemme. Dans le cas où un lemme n'a pas de traits instanciés, ce sera bien entendu la valeur par défaut qui sera choisie, faisant ainsi remonter à la racine (par coindexation) ses propres valeurs, en maintenant leur caractéristique forte ou faible. Si, après une substitution, une de ces valeurs est modifiée, la structure de traits du lemme en sera affectée. La forme sélectionnée sera donc éventuellement modifiée en faveur de celle dont les traits sont unifiables avec ceux de la nouvelle structure.

FIG. 4 – Les régles de reformulation implantées dans PCA.

Ajout d'une unité lexicale
Le déterminant devant un nom commun : *"père"* + *"et"* + *"mère"* ⇒ *"le père et la mère"*
Le pronom sujet : *"vouloir"* + *"dormir"* ⇒ *"je veux dormir"*
Une préposition entre le verbe et le complément :
"il" + *"entrer"* + *"chambre"* ⇒ *"il entre dans la chambre"*
La préposition *"de"* entre deux noms : *"clé"* + *"voiture"* ⇒ *"la clé de la voiture"*
La préposition *"de"* entre le nom et le pronom : *"lit"* + *"je"* ⇒ *"mon lit"* (littéralement *"le lit de moi"*)
Gestion des accords
Accord déterminant-nom : *"fruits"* ⇒ *"les fruits"* ; *"les"* + *"enfant"* ⇒ *"les enfants"*
Accord nom-adjectif : *"beau"* + *"fille"* ⇒ *"la belle fille"*
Accord sujet-attribut : *"elles"* + *"être"* + *"gentil"* ⇒ *"elles sont gentilles"*
Accord sujet-verbe : *"vous"* + *"vouloir"* + *"journal"* ⇒ *"vous voulez le journal"*
Formation de la négation
Positionnement de la négation : *"Pierre"* + *"non"* + *"venir"* ⇒ *"Pierre ne vient pas"*
Déclinaison des pronoms
Nominatif (pronom sujet) : *"ils"* + *"mange"* ⇒ *"ils mangent"*
Accusatif (pronom COD) : *"je"* + *"voir"* + *"elle"* ⇒ *"je la vois"*
Oblique (pronom introduit par une autre préposition que *à*) :
"je" + *"aller"* + *"chez"* + *"tu"* ⇒ *"je vais chez toi"*
Datif (introduit par la préposition *à*) : *"je"* + *"parler"* + *"à"* + *"il"* ⇒ *"je lui parle"*
Gestion des phénomènes linguistiques particuliers
Elision : *"le"* + *"enfant"* ⇒ *"l'enfant"*
Contraction : *"je"* + *"aller"* + *"à"* + *"le"* + *"cinéma"* ⇒ *"je vais au cinéma"*
Concaténation des chiffres en nombre
Concaténation et transformation en déterminant : *"je"* + *"avoir"* + *"1"* + *"5"* + *"an"* ⇒ *"j'ai 15 ans"*
Concaténation des lettres ou groupe de lettres
Si le message est exclusivement composé de lettres : *"b"* + *"on"* + *"j"* + *"ou"* + *"r"* ⇒ *"bonjour"*

2.4 Les règles de reformulation

Les règles de reformulation implantées dans PCA sont issues d'un compromis entre la volonté d'interpréter le maximum de messages iconiques composés tout en demandant un minimum d'informations nécessaires pour caractériser les propriétés syntaxique et sémantique des icônes. De plus, nous avons été amenés à effectuer certains choix d'interprétation pour les situations présentant une ambiguïté sémantique. Nous avons opté pour les règles de reformulation décrites figure 4. Une illustration de message reformulé est présentée figure 5.

Ces règles ont été isolées dans des fichiers ressources externes au programme. Cette architecture nous offre ainsi une certaine souplesse pour faire évoluer la caractérisation et le degré de couverture de l'ensemble des messages interprétables.

FIG. 5 – Une illustration de message iconique reformulé

3 Conclusion

Nous avons présenté dans cet article le module de reformulation implanté dans le mode non verbal de la Plateforme de Communication Alternative. Nous avons choisi de faire porter sur le lexique des informations linguistiques minimales, afin de permettre à l'utilisateur généralement non-expert en linguistique d'enrichir son matériel lexical par ajout de nouveaux items. Ce choix conditionne en pratique les performances du module de reformulation et les règles de génération retenues.

Une cinquantaine de systèmes PCA munis du module de reformulation ont été distribués à ce jour, équipant des particuliers, des professionnels et des structures d'accueil (une version de démonstration est disponible sur le site www.aegys.com). Il est encore trop tôt pour avoir une évalution fiable de l'apport du module de reformulation aux besoins communicationnels des utilisateurs. Néanmoins, les retours qui nous sont parvenus sur l'utilisation du module sont encourageants. Il apparaît que l'oralisation du message reformulé par la synthèse vocale offre à l'utilisateur un contrôle naturel de sa production. Le phénomène est observé aussi bien dans le cadre d'une utilisation de la PCA comme outil de communication que dans le cadre d'une utilisation de la PCA au sein d'un protocole de rééducation.

Références

Abraham M. (2000), “Reconstruction de phrases oralisées à partir d’une écriture pictographique”, Actes de la conférence Handicap 2000, European Journal of Automation, vol. 34, num. 6-7, p. 883-901

Abraham M. (2006), “Altérations de la communication dialogique : Le statut de la langue dans la palliation des troubles de la parole”, Actes de la conférence IFRATH, Handicap 2006 (2006 juin 7-9 : Paris, FRANCE).

Bellengier E., Blache P., Rauzy S. (2004), “PCA : un système d’aide à la communication alternatif évolutif et réversible”, in Actes de la conférence ISAAC 2004, p. 78-85, 6-8 mai 2004, Neuchâtel, Suisse

Bellengier E., Rauzy S., Marty J. (2006), “Système de communication iconique : Reformulation avancée”, Actes de la conférence IFRATH, Handicap 2006 (2006 juin 7-9 : Paris, FRANCE).

Blache P., Rauzy S. (2003) “Linguistic resources and cognitive aspects in alternative communication”, in proceedings of SICS-8, Santiago de Cuba : ISCS. 2003, p. 431-436.

Blache P., Rauzy S. (2004) “Une plateforme de communication alternative”, in Actes des Entretiens Annuels de l’Institut Garches, 26-27 novembre 2004, p. 82-93, Issy-Les-Moulineaux, France

Boissière P., Dours D. (2000) “VITIPI : Un système d’aide à l’écriture basé sur un principe d’auto-apprentissage et adapté à tous les handicaps moteurs”, in Actes de la conférence IFRATH, Handicap 2000 (2000 juin 15-16 : Paris, FRANCE) p. 81-86

Brangier E., Gronier G. (2000) “Conception d’un langage iconique pour grands handicapés moteurs aphasiques”, in Actes de la conférence IFRATH, Handicap 2006 (2000 juin 15-16 : Paris, FRANCE) p. 93-100

Copestake A. (1997) “Augmented and Alternative NLP Techniques for Augmentative and Alternative Communication”, in proceedings of ACL workshop on NLP for Communication Aids, Madrid, Spain. July 12th, 1997.

Joshi87 Joshi A. (1987) “Introduction to Tree Adjoining Grammars”, in A. Manaster Ramer (ed), *The Mathematics of Language*, Benjamins.

Le Pedevic B. (1997) “Prédiction Morphosyntaxique évolutive dans un système d’aide à la saisie de textes pour des personnes handicapées physiques”, Thèse de Doctorat I.R.I.N. (No. ED-82-269)

Maurel D., Fourche B., Briffault S. (2000), “Aider la communication en facilitant la saisie rapide de textes”, in Actes de la conférence IFRATH, Handicap 2000 (2000 juin 15-16 : Paris, FRANCE) p. 87-92

McCoy, K.F., Demasco P.W., Pennington C.A., Luberoff Badman A. (1997) “Some Interface Issues in Developing Intelligent Communication Aids for People with Disabilities” ; Intelligent User Interfaces 1997 : p. 163-170

Pasero R., Sabatier P. (1995), “Guided Sentences Composition : Some problems, solutions, and applications”, in proceedings of NLULP’95, Lisbonne, Portugal, pp 97-110

Vaillant P. (1997), “Interaction entre modalités sémiotiques : de l’icône à la langue”, Thèse de l’Université Paris XI, Orsay, France.

VanRullen T. , Blache P. , Portes C. , Rauzy S. , Maeyhieux J.-F. , Guénot M.-L. ,Balfourier J.-M. , Bellengier E. (2005) “Une plateforme pour l’acquisition, la maintenance et la validation de ressources lexicales”, in actes de TALN, pp. 41-48 (Juin 2005 : Paris, France)

Méthodologie d'annotation des erreurs en production écrite. Principes et résultats préliminaires

Ph. BOISSIÈRE[1], J.-L. BOURAOUI[1], F. VELLA[1], A. LAGARRIGUE[1,2], M. MOJAHID[1], N. VIGOUROUX[1], J.-L. NESPOULOUS[2]
[1] IRIT (Institut de Recherche en Informatique de Toulouse) Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex
[2] OCTOGONE / Laboratoire Jacques Lordat, Université de Toulouse II - Le Mirail Pavillon de la Recherche, 5, allées Antonio-Machado, F-31058 Toulouse Cedex

Résumé. Nous proposons une grille qui cherche à rendre compte, le plus exhaustivement possible, des erreurs survenues dans la production écrite, manuscrite ou clavier, de personnes présentant divers types de handicaps « centraux » ou « périphériques ». L'objectif de ce papier est d'obtenir une identification fine des erreurs survenant pendant la saisie. Celle ci sera ensuite modélisée pour être implémentée dans les systèmes d'assistance à la saisie. Notre grille se décompose en deux parties : la première décrit la nature de l'erreur. La seconde analyse l'erreur et détermine sa conséquence au niveau linguistique. Mettre en œuvre cette grille permet de calculer automatiquement le nombre et le type de fautes pour chaque individu. Le rééducateur (pédagogue, ergothérapeute, etc.) possède ainsi un référentiel lui indiquant les faiblesses langagières du patient dans un but de rééducation. Nous décrivons notre méthodologie et donnons les premiers résultats obtenus à partir d'écrits d'adolescents IMC.

Abstract. We propose a grid which seeks to make an exhaustive explanation of the errors occurring during the writing production, handwriting or keyboard, for people presenting various types of "central" or "peripheral" handicaps… The aim of this paper is to obtain a subtle modeling of the errors occurring during a keyboarding. The result will then be introduced in assistance systems to the keyboarding.. This grid is made up of two parts: the first describes the nature of the error. The second analyzes the error and establishes its linguistic consequence. To implement this grid in a spreadsheet makes it possible to automatically calculate the number and the type of faults for each individual. Thus the reeducator (pedagogue, occupational therapist, etc.) has a tool which indicated the linguistic weaknesses of the patient and with which he can try to rehabilitate it. We give the first results obtained from IMC teenagers' writings.

Mots-clés : analyse et typologie d'erreurs langagières, modélisation linguistique, handicaps langagiers, assistance automatique à la saisie de textes.

Keywords: analysis and typology of textual errors, automatic assistance to texts input.

1. Introduction

Errare humanum est. Cet adage se vérifie partout, y compris évidemment pour la communication et à l'expression. Dans le contexte du dialogue oral, il est fréquent que les erreurs soient repérées. Or, quand ce n'est pas le cas, les conséquences peuvent poser de gros problèmes. Il faut par conséquent maximiser l'identification et la compréhension des erreurs d'ordre linguistique, et minimiser leurs conséquences. C'est un gros travail puisqu'il concerne tous les modalités de communication langagière, notamment les deux principaux : l'oral et l'écrit. De plus, il existe différents objectifs et méthodes pour aborder les erreurs à tous les niveaux de production propres à chaque modalité, même s'ils peuvent s'harmoniser.

Le travail présenté ici suit une démarche d'ordre neuro-psycholinguistique pour l'écrit. Il trouve son origine première dans les travaux menés par l'un de nous (Jean-Luc Nespoulous), en collaboration avec André Roch Lecours (1979) sur les productions écrites déviantes des patients aphasiques. Il est également complété, sur tel ou tel point, par certains éléments en provenance des travaux sur l'écriture de Nina Catach (1980) et de son équipe.

Le thème central de l'article est de proposer une grille d'interprétation et d'annotation des erreurs à l'écrit. L'objectif de cette grille est de rendre compte, de la manière la plus exhaustive possible, des erreurs survenues dans la production écrite, manuscrite et sur clavier de personnes présentant divers types de handicaps « centraux » ou « périphériques ».

Cet article s'inscrit dans le projet ESACIMC[1]. L'Infirmité Motrice Cérébrale (IMC) est due à une mauvaise oxygénation des cellules cérébrales dans la période anté ou post natale. Une partie du cerveau est lésé, entraînant des problèmes moteurs qui peuvent parfois troubler la communication orale voire même atteindre les aires du langage. La combinaison de ces problèmes implique forcément des difficultés d'écriture qui, même avec l'outil informatique, est très lente et parfois perturbée orthographiquement. Par exemple, la phrase suivante a été reconstruite à partir de fautes trouvées dans un corpus d'adolescent IMC (18 ans) : * « les médcin on sinier poyr mon deuxièm fauteuil electric. c'est bein. j'ai merai etre en musur daprand un maitié pour pas allé au foyé,». On peut espérer qu'un correcteur orthographique pourra en réparer certaines, cependant, aussi performant soit-il, lorsque les mots possèdent plusieurs fautes, la tâche parait impossible[2]. D'où l'idée de remédier –en partie – à ce problème en commençant par analyser le plus finement possible les fautes commises par ces sujets de façon à les répertorier et à comptabiliser leurs fréquences. Nous pourrons ensuite modéliser les fautes commises puis, dans une dernière étape, automatiser la correction. Tel est notre objectif à long terme, certes ambitieux, dont nous vous présentons les prémisses.
Les corpus utilisés ont été écrits dans ce contexte, mais les conclusions qui sont tirées sont généralisables.

Nous présentons d'abord les principaux enjeux et problèmes liés à notre étude. Nous passons ensuite en revue les différentes catégories d'erreurs et de perturbations qui leur sont associées. Nous montrons notamment que lors de la saisie par clavier, les erreurs peuvent avoir une motivation phonétique, morphémique, voire spatiale[3]. Nous donnons enfin quelques résultats.

[1] Evaluation qualitative de Systèmes d'Aide à la Communication pour les Infirmes Moteurs Cérébraux
[2] A l'exclusion des fautes d'accord et de ponctuation, celui de Microsoft Word 2003 en détecte 12 et en corrige 6.
[3] La méthodologie d'annotation et ses bases théoriques sont décrites bien plus en profondeur dans (Bouraoui et al., 2007)

2. Réflexions et analyses préliminaires

2.1. De la nécessité (et de la difficulté) d'une analyse « multi-niveaux » des erreurs de la production écrite

Compte tenu de l'existence de différents niveaux d'organisation de la structure des langues naturelles, il faut rendre compte de l'ensemble des erreurs susceptibles de survenir à chacun de ces niveaux : littéral, graphémique, lexical, morphologique, syntaxique … et portant sur des entités linguistiques allant de la « lettre » à la « phrase » et au « texte ».

Rares sont finalement les erreurs qui (a) **ne se situent qu'à un seul niveau** et (b) n'ont pas **d'impact (même indirects) à d'autres niveaux**. Ainsi, telle omission « locale » d'une préposition entraîne l'agrammaticalité de la phrase dans laquelle elle intervient (ex. : « Il a posé l'assiette X**XX** la table »). Pareillement, une erreur qui pourrait n'être qu'orthographique peut entraîner, secondairement, une violation morphologique (ex. : « il mangeais »).

Nous adoptons une démarche « multi-niveaux » qui demande d'octroyer à une même erreur « superficielle » plusieurs étiquettes. Dans le premier exemple ci-dessus, nous on sommes amenés à étiqueter à un premier niveau, l'omission de morphème grammatical en tant que telle, avant d'ajouter une deuxième étiquette, au plan syntaxique cette fois (« agrammatisme »). Le sujet n'a pas commis plusieurs erreurs. Cela veut dire qu'en commettant une erreur « locale » à tel ou tel endroit du message, il a entraîné plusieurs violations aux conditions de bonne formation des énoncés (ici écrits). Ceci nous conduit à différencier, (a) des « erreurs locales à impact (simplement) local » et (b) des « erreurs locales à effets secondaires », ces dernières présentant un degré de gravité plus important, susceptible de perturber de façon massive l'échange d'informations. On aura bien compris que ce point complique passablement l'analyse dont il est ici question.

2.2. Principaux problèmes rencontrés lors de l'analyse de l'écrit

En dehors des difficultés mentionnées ci-dessus, il y a également d'autres sources de problèmes d'analyse de l'écrit. Nous en présentons les principales :

- Rapport avec l'oral : il est inenvisageable d'analyser l'écrit sans prendre en considération l'oral. L'écrit n'est en effet qu'une transcription de l'oral qui a été acquis le plus souvent en premier. Dès lors, bon nombre d'erreurs dans la production écrite sont influencées (« contaminées ») par la nature des représentations orales, souvent « co-activées » lorsque le sujet entreprend sa tâche d'écriture ! L'existence d'homophones non homographes constitue par conséquent un problème majeur dont une grille d'analyse complète doit pouvoir rendre compte ;

- La chronométrie : les paramètres temporels sont de nature à permettre une analyse plus fine de la production écrite. Par exemple, l'arrêt du scripteur entre l'écriture du radical et celle de la désinence d'un verbe indique vraisemblablement que le sujet n'est pas à l'aise dans sa gestion de la morphologie flexionnelle verbale ! Il semble notamment important de prendre en considération, (a) les éventuelles mauvaises segmentations de mots, (b) les problèmes majeurs de ponctuation, voire (c) les tentatives d'autocorrection. Ces dernières, si elles ne sont pas systématiquement relevées peuvent conduire à des erreurs de diagnostic ;

- La configuration spatiale des lettres sur un clavier : En écriture sur clavier, à ces erreurs (toujours possibles) s'ajoutent celles qui peuvent émaner de la proximité spatiale de certaines lettres, et ce, même si « lettres substituantes » et « lettres substituées » n'ont rien en commun dans le système alphabétique. Par exemple, les touches correspondant aux lettres Z, E, R, S, D, F sont toutes situées sur la même zone d'un clavier AZERTY. Mais elles n'ont aucun point commun en termes de graphie ou de sonorité. Ce point risque d'être crucial pour divers types de populations pathologiques présentant des problèmes moteurs importants (ainsi d'ailleurs, dans une moindre mesure, que pour tout autre catégorie de public). On peut envisager de modéliser ce type d'erreurs en attribuant une « pondération » aux lettres faisant l'objet d'une erreur, en fonction de leur proximité plus ou moins grande sur le clavier ;

- Erreurs VS stratégies : il arrive que certains phénomènes erronés ne soient pas la conséquence directe d'un « déficit », mais plutôt, la mise en œuvre de stratégies (plus ou moins volontaires) susceptibles de faciliter la tâche à l'émetteur ; les écrits SMS en fournissent de bons exemples : « g » pour « j'ai »…). La systématicité d'une erreur peut aussi indiquer la mise en place d'une « stratégie », d'où les difficultés déjà mentionnées ! En TALN (Traitement automatique du Langage Naturel), cette distinction est fondamentale puisque c'est à partir de phénomènes linguistiques invariants que l'on peut établir des modèles et des algorithmes.

3. Typologie des erreurs

Nous présentons d'abord les différents niveaux auxquels une analyse des erreurs intervenant à l'écrit (avec un focus sur la saisie clavier) peut être menée. Nous nous livrons ensuite à une catégorisation détaillée des erreurs pouvant survenir au niveau choisi.

3.1. Approche macroscopique

A l'écrit, les différents niveaux auxquels interviennent les erreurs sont au nombre de 5. On peut les hiérarchiser, du plus global au plus particulier, de la manière suivante :

1. Recours (ou non) à de la MFM[4] (au plan spatial essentiellement) ;

2. Problèmes de segmentation en phrases (i.e. énoncés phrases, même si celles-ci contiennent des erreurs) VS énoncés non-phrases (i.e. « style télégraphique ») ;

3. Problèmes au niveau de l'unité lexicale : il peut s'agir aussi bien l'unité lexicale dans sa globalité que de ses entités constituantes (lettres, morphèmes, etc…) ;

4. Problèmes de gestion de la ponctuation : point VS virgule, au niveau phrastique et intra-phrastique ;

5. Problèmes de gestion des blancs inter-mots, voire intra-mots, ces derniers sont appelés « erreurs logogrammiques » par (Catach, 1980).

[4] Mise en Forme Matérielle du texte. Pour résumer, cette théorie (Virbel, 1989) postule notamment la prise en compte de la mise en forme du texte écrit (indentation, mise en gras, etc.) dans la transmission du sens. Par exemple la mise en italique d'un mot indique un focus particulier porté sur celui-ci. Mentionnons les travaux de (Luc, 2000) sur l'architecture des énumérations et la typographie des paragraphes, qui fournit des éléments de réflexion pour le niveau 1.

Dans ce travail, nous ne nous sommes penchés que sur le niveau 3 pour plusieurs raisons. D'une part, il s'agit du niveau le plus facile à modéliser et pour lequel on peut implémenter des améliorations dans des logiciels d'assistance. D'autre part, on retrouve des analogies avec les productions orales. Enfin, il a fait l'objet de nombreuses études et expérimentations, notamment psycholinguistiques, sur lesquelles nous pouvons nous baser. Il est important de signaler ici qu'à ce niveau, nous n'avons pas pris en compte les erreurs relatives aux noms propres, majuscules, et à certains usages de langage très soutenus[5].

On notera que ce niveau couvre un large spectre. Il nécessite donc une sous catégorisation fine, que nous décrivons dans ce qui suit.

3.2. Approche analytique

3.2.1. Types d'erreurs

Le but est de permettre une description aussi précise que possible des erreurs, du niveau le plus concret jusqu'au plus abstrait. Pour cela, on utilise une hiérarchie de catégories. Celle-ci est représentée dans la figure 1, que nous explicitons immédiatement après.

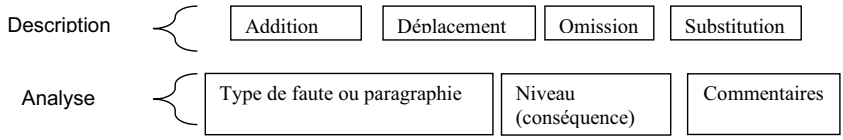


Figure 1 : Hiérarchie des catégories d'erreurs.

Le premier niveau, « *description* », correspond à la manifestation de la faute. En effet, toute faute appartient systématiquement à l'une et une seule de ces catégories : *Addition*, *Déplacement*, *Omission*, *Substitution*. Nous nous y réfèrerons via l'acronyme *ADOS*. Cette catégorisation a été modélisée par Levenshtein (1966), et a donné lieu à de nombreux développements et applications dans diverses disciplines liées au TALN.

Le deuxième niveau, « *analyse* », correspond à une tâche plus délicate. Il s'agit de catégoriser plus finement l'erreur, de situer le niveau qu'elle affecte, et d'avancer des hypothèses motivant son apparition. Dans la partie suivante, nous décrivons chacune de ces étapes.

Quand une seule erreur de l'un de ces types apparaît dans un énoncé, l'analyse ne pose pas de problème. Cependant, si plusieurs erreurs de l'un ou l'autre de ces types apparaissent (surtout dans le cas des omissions), l'analyse devient plus complexe, *a fortiori* s'il s'agit de l'omission de plusieurs mots grammaticaux. Dans cas, il vaudra mieux recourir à l'étiquette d'« agrammatisme » pour caractériser l'énoncé en question (sans chercher à quantifier le nombre d'omissions de morphèmes… ce qui de plus, s'avèrerait quasiment impossible)6.

3.2.2. Types d'unités linguistiques perturbées (et leur interprétation)

S'agissant du langage écrit, les unités suivantes sont pertinentes et sont toutes susceptibles d'être « malmenées » en situation de production écrite.

[5] Pour plus de détails, cf (Bouraoui et al., 2007), III.3.
[6] Alors que les omissions, ici ou là, de lettres peuvent, elles, être aisément quantifiées.

Reprenons les deux niveaux évoqués dans la figure 1, avec pour chacun les différentes catégories correspondantes :

Niveau « Description » *:* Les différentes catégories référencées *ADOS* sont assez explicites, et des exemples sont donnés dans la section suivante. Mais il faut apporter deux précisions.

D'une part, l'unité pouvant faire l'objet d'un des phénomènes catégorisés est en général la lettre (ou le caractère du clavier), mais selon les circonstances, il pourra s'agir de plusieurs caractères, ou encore d'un ou plusieurs phonèmes. On notera qu'en français, lettre et phonème ne sont pas systématiquement dans une relation bijective. En effet, des phonèmes peuvent être réalisés orthographiquement par plusieurs lettres pour aboutir à un « graphème ». Un exemple typique est le phonème /ã/, qui s'écrit « empts » dans « exempt » ou le /a/ de « femme ».

D'autre part, les erreurs concernant les diacritiques font l'objet d'un traitement particulier quand on utilise un clavier d'ordinateur pour produire le texte. Nous ferons en effet la distinction entre d'un côté les lettres accentuées présentes « d'un bloc » sur une seule touche de clavier (en français : ç, é, è, à, ù), et de l'autre celles formées par la combinaison de 2 touches (en français, diacritique ^ ou ¨, et une voyelle). Dans le premier cas, une seule faute sera comptée. Dans le second, on pourra compter 1 à 2 fautes : une pour le diacritique, et une, éventuellement, pour la lettre accentuée.

On l'a vu, pendant l'écriture, une lettre ou un groupe de lettres peuvent être altérés ou écrits d'une autre façon avec une autre graphie. Ainsi, peuvent se produire des *paragraphies* provenant d'origines diverses et dont nous analyserons les conséquences.

Ici, la classification est faite selon l'unité linguistique perturbée :

- La « lettre » : une erreur de ce type sera appelée « *paragraphie littérale* » (**PL**)

  Ex[7] : TORNADE ➔ *TORNADRE addition (avec persévération)
  CULTIVATEUR ➔ *CURVILATEUR déplacement
  CHERCHER ➔ *CHECHER omission
  FOURNIR ➔ *FOURFIR substitution (avec persévération)

  C'est à cette première catégorie d'erreurs que viendront se rajouter les Paragraphies Littérales engendrées par proximité des touches sur le Clavier (**PLC**). Pour une prise en compte des PLC dans un système d'assistance à l'écriture, cf. (Boissière & Dours 1996, p.170).

  Concernant la catégorie des PLC, il est important de préciser qu'on ne doit l'employer seulement que pour les erreurs d'addition, ou de substitution entre une lettre cible et un périmètre situé immédiatement autour[8] (faute de frappe).

  Appartiennent également à cette catégorie, les erreurs portant sur les diacritiques (accents) : omissions VS substitutions (ces dernières venant modifier la lecture à haute voix des mots écrits de manière erronée. Ex : « fenêtre » ➔ « * fénêtre »).

- Le « graphème » : une erreur de ce type sera appelée *« paragraphie graphémique »* (**PG**).

[7] Exemples tirés de (Lecours, A.R et al.1979).
[8] En attendant qu'éventuellement des études plus poussées élargissent le champ d'application des PLC.

Il s'agira essentiellement de substitutions. Le « graphème » est l'équivalent (ortho)graphique d'un phonème. Une « lettre » peut correspondre à un graphème mais souvent, surtout dans une langue comme le français, un graphème nécessite l'emploi de plusieurs lettres (parfois nombreuses) comme nous le signalons plus haut.

Cette catégorie est donc surtout utile pour qualifier les erreurs de production écrite de *phonèmes hétérographes* : le sujet utilise une variante graphémique erronée (qui, néanmoins, permet de renvoyer au bon phonème à l'oral).

Ex : BIBLIOTHEQUE ➔ *BIBLIOTEC
FRANÇAIS ➔ *FRANÇAIT
COMMENCEMENT ➔ *COMANSEMENT

- Le « morphème » : une erreur de ce type sera appelée *« paragraphie morphémique »* (**PM**).

  Il s'agit des « erreurs morphogrammiques » de Catach. Dans sa forme la plus simple à analyser, il s'agira d'une substitution de morphèmes (le plus souvent flexionnels) : « il chantait » ➔ « * il chantais ».

  Rentrent également dans cette catégorie les erreurs d'accords morphologiques, les omissions ou additions de morphèmes grammaticaux et les erreurs de préfixes (par exemple IMPOSSIBLE ➔ * INPOSSIBLE).

  Il y a enfin une autre catégorie, celle des *« Perturbations Morphémiques »* (**PeM**). Elle se divise en deux sous-catégories :

  - *« Perturbation Morphémique fusion»* (**PeMf**) : omission d'une lettre (caractère) séparant deux substantifs (espace, trait d'union, apostrophe), et aboutissant ainsi à la réunion de ceux-ci. Ex TRAIT D'UNION: ➔ * TRAIT DUNION ;

  - *« Perturbation Morphémique segmentation»* (**PeMs**) : addition d'une lettre (caractère) séparant (espace, trait d'union, apostrophe) à l'intérieur d'un substantif, aboutissant ainsi à la segmentation de ce dernier en deux unités. Ex : PROPOSER ➔ * PROP OSER.

Niveau « analyse » : les différents niveaux pouvant être affectés sont les niveaux orthographique, phonologique, morphologique et syntaxique.

En règle générale, les fautes relevant de la PL affectent seulement le niveau orthographique et/ou phonologique ; les PG et les PM peuvent affecter, selon les cas, les 4 niveaux. Voici quelques indications supplémentaires pour chaque catégorie (hors orthographe) :

- *Phonologique* : pour déterminer si une erreur concerne le niveau phonologique, il suffit de comparer la représentation phonétique du mot erroné avec celle du mot cible, par exemple en lisant à haute voix. Si les deux représentations sont identiques, alors on ne parle pas « d'erreur à conséquence phonologique ». L'interprétation peut différer selon les variantes locales, comme la prononciation ou non du e final.

- *Morphologique* : à ne pas confondre les erreurs morphologiques et syntaxiques. Les erreurs d'ordre morphologiques, comme leur nom l'indique, concernent uniquement la morphologie grammaticale du mot. Il peut s'agir par exemple d'une faute de flexion (conjugaison, genre, nombre …).

- *Syntaxique* : le terme de « syntaxe » est pris ici au sens de la position et de la fonction des mots dans l'énoncé. Donc, on ne parlera d'erreur de syntaxe que dans le cas d'erreurs à ce niveau. Par exemple, l'omission d'un substantif, ou son déplacement à une autre position que celle qui lui est assignée dans la langue.

Précisons enfin que nous avons été amenés à définir des règles pour déterminer ce qui constitue une erreur ou pas. Nous n'avons pas voulu centrer notre publication sur ce point ; le lecteur intéressé se reportera à (Bouraoui, 2007).

4. Statistiques préliminaires

Nous avons extrait d'un corpus de 472 phrases fournies par le centre KERPAPE[9] dans le cadre du projet ESACIMC[10], un ensemble de 13 phrases provenant d'un même sujet. Ces phrases ont été analysées selon notre méthodologie. L'annotation se faisant sous Excel, nous pouvons obtenir directement à l'aide de macros les résultats suivants. Dans chaque figure, les nombres en gras situés à gauche des pourcentages correspondent aux nombres d'occurrences.

4.1. Description

Sur la figure 2, nous constatons que près de la moitié des fautes sont des omissions (46 %) suivies des substitutions (39 %). La lecture de la figure 3 montre que seulement 1 % de ces erreurs sont des fautes de frappes (ce qui peut surprendre pour un sujet IMC). Les paragraphies littérales (27 %), graphémiques (34 %) et morphémiques (22 %) représentent à elles seules (83 %) des erreurs. Cela indique probablement des difficultés orthographiques importantes.

[9] Centre Mutualiste de Rééducation et de Réadaptation Fonctionnelles de Kerpape (www.kerpape.mutualite56.fr)
[10] http://www.irit.fr/ESACIMC/

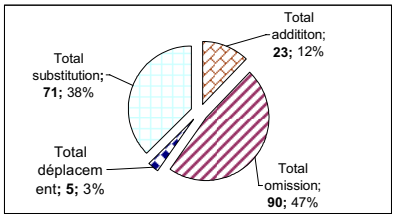


Figure 2 : Répartition des fautes.

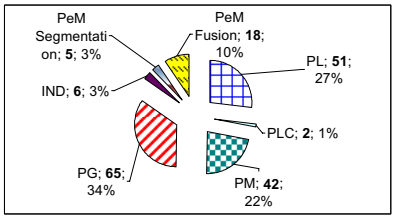


Figure 3 : Distribution des types de fautes.

4.2. Analyse

Au niveau de l'analyse, nous constatons (Cf. Figure 4) qu'effectivement, 51 % des fautes sont à conséquence orthographiques, et 26 % à conséquence morphologiques. Seulement 3 % sont des fautes de syntaxe (oubli de mot) et 20 % des fautes entraîneraient une mauvaise lecture par une synthèse vocale. Nous envisageons de mettre ces résultats en corrélation avec les tableaux cliniques des sujets (pathologie, degré de qualification de la pathologie, parcours scolaire, etc.).

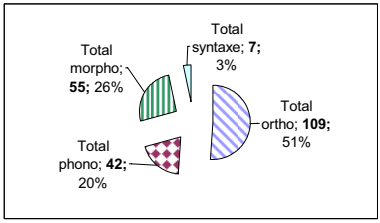


Figure 4 : Distribution des niveaux affectés.

5. Conclusion et perspectives

Le travail que nous avons présenté n'en est encore qu'à ses débuts, il doit être complété sur l'ensemble du corpus de Kerpape. Nous avons fait annoter un corpus d'une douzaine de phrases par deux annotateurs différents, à partir des règles décrites dans ce document, et détaillées dans (Bouraoui et al., 2007). A court terme, nous envisageons de mener une campagne d'annotation à plus grande échelle, sur plus de corpus (corpus thématiques différents, sujets avec des handicaps langagiers différents, etc.) et avec plus d'annotateurs pour définir le coefficient de Kappa (Carletta 1996). Celui ci nous permettra d'évaluer de manière fiable et objective la stabilité et la robustesse de notre grille d'annotation. Nous comptons également mieux mettre en place un corpus « artificiel », conçu à partir de diverses tâches de production écrite identiques pour tous les sujets, éventuellement réutilisées chez les mêmes sujets à différents moments (étude longitudinale). Ce corpus serait complémentaire des corpus actuels.

Enfin, à plus long terme, nous envisageons de modéliser ces erreurs dans les connaissances linguistiques du système d'assistance à la saisie VITIPI. (Boissière, 1996).

Remerciements

Cette étude a été réalisée grâce aux financements de l'APETREIMC et de la Fondation Motrice. Les auteurs expriment également leur reconnaissance à nos partenaires du Centre Mutualiste de Rééducation et de Réadaptation Fonctionnelles de Kerpape qui nous ont gracieusement fait parvenir les textes écrits par certains de leurs patients. Que ces derniers en soient ici sincèrement remerciés.

Références

BAUDOT J.A. (1968) *Information, redondance et répartition des lettres et des phonèmes en français*, Rapport, Université de Montréal, mars 1968.

BOISSIÈRE PH., DOURS D. (1996) "VITIPI : Versatile Interpretation of Text Input by Persons with Impairments". In *5th ICCHP (International Conference on Computers for Handicapped Persons)*. pp.165-172, Linz July 1996.

BOURAOUI J.-L., BOISSIÈRE PH., VELLA F., LAGARRIGUE A., MOJAHID M., LAUR D., VIGOUROUX N., NESPOULOUS J.-L. (2007) *Prolégomènes à l'étude des erreurs en production écrite - Propositions en vue de la mise au point d'une grille d'analyse*, Rapport IRIT/RR—2007-7-FR, Mars 2007.

CARLETTA J. (1996) "Assessing agreement on classification tasks: the kappa statistic", Computational Linguistics, 22 (2):249-254, 1996.

CATACH N. (1980) *L'enseignement de l'orthographe*, Paris, Nathan.

LECOURS A. R. (1966) "Serial order in writing – a study of misspelled words in "developmental dysgraphia"", *Neuropsychologia*, Vol.4, pp. 221-241.

LECOURS A. R., DELOCHE G., LHERMITTE F. (1973) Paraphasies phonémiques – description et simulation sur ordinateur –, in *Colloque INRIA-Informatique Médicale*, pp.311-351, Rocquencourt.

LECOURS A. R., LHERMITTE F. (1969) Phonemic paraphasias: linguistic structures and tentative hypotheses, *Cortex*, 5, pp.193-228.

LECOURS A. R., LHERMITTE F. ET AL. (19) *L'aphasie*, Paris, Flammarion.

LECOURS, A.R., DORDAIN, G., NESPOULOUS, J-L. & LHERMITTE, F. (1979) « Le vocabulaire de la neurolinguistique », in A.R. Lecours & F. Lhermitte (Eds) *L'aphasie*, Paris, Flammarion.

LECOURS, A.R., NESPOULOUS, J-L (1982) « Biologie de l'écriture », *Etudes Françaises*,18/1, 33-45, Les Presses de l'Université e Montréal.

LEVENSHTEIN V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals, *Cyber. Contr. Theory*,10 (8), pp. 707-710.

LUC C. (2000) *Représentation et composition des structures visuelles et rhétoriques du texte. Application à la génération de textes formatés.* Thèse de doctorat, Université Paul Sabatier, novembre 2000.

MOUNIN G. (1970) *Introduction à la sémiologie*, Paris, Editions de Minuit.

NESPOULOUS, J-L., LECOURS, A.R. (1982) « Les troubles de l'écriture dans l'aphasie », *Etudes Françaises*, 18/1, pp. 47-59, Les Presses de l'Université de Montréal.

NESPOULOUS, J-L. & VIRBEL, J. (2004) Apport de l'étude des handicaps langagiers à la connaissance du langage humain, *Revue Parole*, N°29-30, pp. 5-42.

VIRBEL, J. (1989). "The contribution of linguistic knowledge to the interpretation of text structure". Dans Andre, J., Quint, V. et Furuta, R., (Eds) *Structured Documents*, pages 161–181. Cambridge University Press.

Système Sibylle d'aide à la communication pour personnes handicapées : modèle linguistique et interface utilisateur

Tonio WANDMACHER[1,2], Nicolas BÉCHET[1,4], Zaara BARHOUMI[3],
Franck POIRIER[3], Jean-Yves ANTOINE[1]
[1] Université François Rabelais de Tours – LI, IUP Blois, France
[2] Universität Osnabrück – Université2, Adresse2
[3] Université Européenne de Bretagne – VALORIA, UBS Vannes, France
[4] Université Montpellier II – LIRMM, Montpellier, France
{jean-yves.antoine,tonio.wandmacher}@univ-tours.fr
franck.poirier@univ-ubs.fr, nicolas.bechet@lirmm.fr

Résumé. Cet article est consacré à une description complète du système SIBYLLE d'aide à la communication qui est développé conjointement par les laboratoires LI et VALORIA. Cet article décrit à la fois le module de prédiction linguistique qui repose sur des modèles de langage avancés, et l'interface utilisateur qui a été développée pour prendre en compte les besoins réels des utilisateurs tels qu'ils ont été définis avec le centre de rééducation fonctionnelle de Kerpape. Les performances du système sont présentées, de même que l'intégration de SIBYLLE dans le projet francophone ESACIMC d'analyse des usages réels des communicateurs.

Abstract. This paper describes the AAC system SIBYLLE, which is developed jointly by the LI and VALORIA laboratories. It presents the word prediction module of SIBYLLE, which is based on an advanced language model, but also its user interface, which follows the recommendations of the centre of functional rehabilitation of Kerpape. The most significant performances of the system are presented. Finally, we describe the involvement of SIBYLLE in the ESACIMC project.

Mots-clés : handicap et TALN, systèmes d'aide à la communication, prédiction de mots, interaction homme-machine.

Keywords: NLP and handicap, Alternative and Augmentative Communication, word prediction, computer-human interaction.

1 Aide à la communication pour personnes handicapées

Les communicateurs, ou systèmes de communication palliative (AAC pour *Alternative and Augmentative Communication* en anglais) ont pour objectif de restaurer les capacités de communication de personnes souffrant d'un handicap moteur très sévère (Infirmités Motrices Cérébrales, Scléroses Latérales Amyotrophiques, syndrome d'enfermement,…) se traduisant par une tétraplégie ou une athétose accompagnée d'une perte de l'usage de la parole. La communication est alors privée de son support oral habituel, de même que les capacités très limitées de contrôle physique de l'environnement par la personne handicapée empêchent toute saisie directe de message sur un clavier d'ordinateur.

Ces systèmes repose sur l'écriture de phrases à l'aide d'un clavier virtuel affiché à l'écran. Dans le cadre de clavier à défilement linéaire, un curseur se déplacer caractère par caractère, le long du clavier. L'intervention de la personne handicapée se limite à la désignation des symboles lorsque le curseur est sur la touche ou le caractère désiré. Cette sélection est réalisée à l'aide d'un dispositif physique qui remplace le périphérique d'entrée de l'ordinateur. Cette interface matérielle dépend des capacités motrices de l'utilisateur. Il peut s'agir d'un joystick, d'une commande oculaire, d'une commande par souffle, d'un simple bouton poussoir, etc. Une caractéristique importante est le degré de liberté qu'elle permet pour manipuler l'ordinateur. Le plus souvent, le patient n'a plus que la possibilité de réaliser l'équivalent d'un simple clic (commande de l'environnement de type « tout ou rien »). Une fois le message saisi, il peut être vocalisé par l'intermédiaire d'une synthèse de parole artificielle (*text-to-speech synthesis*).

Le problème majeur des systèmes de communication assistée est la lenteur de la composition des messages. La tâche de saisie est généralement longue (1 à 5 mots par minute en moyenne) et fatigante (BérarD 2004 ; Vella & Vigouroux 2007) pour les sujets. Pour accélérer la saisie, deux approches complémentaires sont envisageables. La première vise à optimiser la sélection sur le clavier simulé en faisant en sorte que le curseur défilant arrive au plus vite sur le caractère recherché. La seconde consiste à limiter le nombre de saisies en prédisant les mots qui peuvent survenir à la suite de ceux qui ont déjà été saisis. Plusieurs méthodes peuvent être utilisées pour réaliser ces optimisations. SIBYLLE adopte une démarche ascendante (partant des données déjà saisies) qui repose sur l'utilisation de modèles markoviens de langage.

2 Modèles de langage pour l'optimisation de la saisie de message

2.1 SIBYLETTRE : optimiser le temps d'accès à un caractère

Plusieurs approches peuvent être envisagées pour accélérer l'accès au symbole recherché. Certaines concernent directement la disposition des touches sur le clavier (Cantegrit, Toulotte 2001; Vella, Vigouroux 2007). Ici, nous ne considérerons que les techniques d'optimisation de nature linguistique, en nous plaçant dans le cas d'un curseur à défilement linéaire (pas de balayage ligne-colonne). Un moyen d'optimiser la sélection est alors de refondre dynamiquement la disposition du clavier après chaque saisie, afin que les symboles les plus probables compte tenu des lettres précédentes soient balayés en premier (figure 1).

Cette réorganisation est souvent basée sur la consultation d'un dictionnaire fréquentiel: on propose la lettre qui conduira au mot le plus fréquent du dictionnaire à partir de l'amorce déjà saisie. Cette technique, très efficace, pose cependant le problème des mots hors vocabulaire mais aussi des fautes de saisie ou d'orthographe, très fréquentes avec les systèmes AAC. Dans ces situations, le système est totalement perdu et l'optimisation est inopérante.

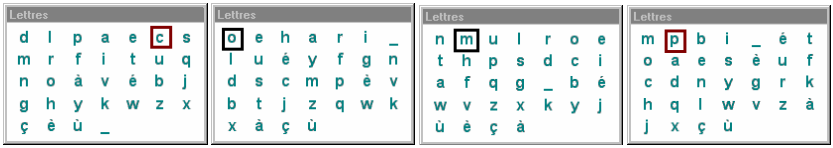


Figure 1 : Exemple de réorganisation dynamique sur le début de mot COMP…TER

C'est pourquoi nous avons choisi d'adopter dans le système SIBYLLE une analyse infra-lexicale basée sur un modèle N-gram au niveau des lettres. Les lettres du clavier sont

réorganisées en fonction de l'estimation de leur probabilité d'occurrence compte tenu des quatre dernières lettres saisies, espace et caractères de ponctuation compris : $P(c_i) \approx P(c_i \mid c_{i-1}, \ldots, c_{i-4})$.

Mode de défilement	Linéaire Azerty	Ligne/Colonne Azerty	Sibylettre Français	Sibylettre-allemand	
				standard	avec majuscules
défilements / caractère	33	9	2,9	3,0	3,7

Tableau 1 : Nombre moyen de défilements nécessaire pour la saisie d'un caractère. Test réalisé sur des extraits de corpus journalistiques de 50 000 mots.

Les probabilités sont estimées sur un corpus représentatif (corpus journalistique quelle que soit la langue considérée). Le modèle donne des résultats aussi satisfaisants qu'une consultation de lexique, avec l'avantage de subir une dégradation limitée de performance en cas d'erreur de saisie ou de faute d'orthographe. Ainsi, quelle que soit la langue considérée (voir tableau 1), le caractère attendu se trouve en moyenne dans les 3 ou 4 premiers symboles proposés. Ces performances correspondent à un gain de performances très sensible par rapport à un balayage ligne-colonne (9 défilements par caractère en moyenne) ou un clavier Azerty statique (33).

2.2 SIBYMOT : éviter les saisies grâce à la prédiction de mots

Les performances du module SIBYLETTRES sont proches du minimum théorique de défilements par caractère que l'on peut espérer (2,7 pour le français) sans considération de l'organisation des mots dans l'énoncé. C'est précisément à cette analyse du langage que procède le module SIBYMOT qui prédit les mots à venir afin d'éviter à l'utilisateur les saisies correspondantes.

SIBYMOT adopte une approche ascendante qui consiste à prédire le mot courant en fonction des précédents déjà saisis et, éventuellement de ses premières lettres. Après chaque saisie, une liste de prédictions lexicales est présentée dans un sous-clavier spécifique (voir figure 2). Si l'utilisateur retient une de ces propositions (la sélection d'une touche spécifique permet le basculement du défilement d'un sous-clavier à un autre), le texte est automatiquement complété, ce qui évite la saisie des dernières lettres du mot.

L'efficacité de ces techniques de prédiction est évaluée par le taux d'économie de saisies ou KSR (*keystroke saving rate*) : $KSR = (1 - kp/ka) \cdot 100$ où *kp* et *ka* représentent respectivement le nombre d'appuis sur le dispositif d'entrée avec et sans prédiction.

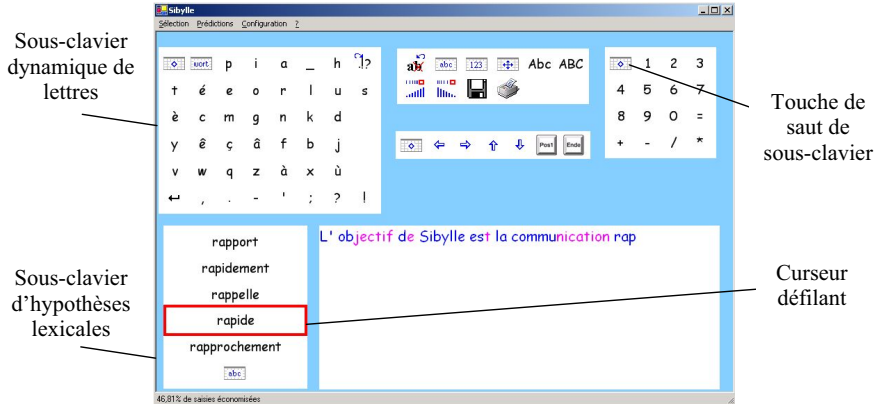


Figure 2 : Interface de la version 2.4 du système SIBYLLE, avec prédiction de lettres et de mots

Dans SIBYMOT, la prédiction repose sur un modèle stochastique de langage auquel ont été apportés plusieurs raffinements. Le modèle de base est un quadrigramme de mots, c'est-à-dire que l'on estime la probabilité d'apparition d'un mot compte tenu de ses trois prédécesseurs : $P(w_i) \approx P(w_i \mid w_{i-1}, \ldots, w_{i-3})$. Nous avons développé des modèles de langage pour le français, l'allemand et l'anglais. Ils ont été entraînés sur des corpus journalistiques de taille comparables (50 à 100 millions de mots) à l'aide du *toolkit* du *SRI* (Stolcke, 2002) employé avec un lissage de type Kneser-Ney (Goodman, 2001) et la méthode de *pruning* proposée par (Stolcke, 1998).

Ce modèle présente déjà d'excellents résultats, puisqu'il permet d'économiser en moyenne plus de la moitié des saisies sur un corpus de test de même registre de langue (KSR supérieure à 50% sur des articles de journaux). Cependant, plusieurs problèmes se posent pour son application directe à l'aide à la communication pour personnes handicapées :

- *Adaptation à l'utilisateur* – Quelque soit l'objectif de la communication, il est clair que le style de l'utilisateur sera différent de celui d'un journaliste. Il faut donc adapter la prédiction pour intégrer le vocabulaire propre à l'utilisateur et son style de langage.

- *Adaptation au registre et au thème du discours* – Chaque type de communication (dialogue oral, courriel, courrier officiel, roman, …) répond à un registre (Biber, 1993) particulier vers lequel l'utilisateur peut basculer à tout moment. Il est donc intéressant d'adapter dynamiquement la prédiction au registre courant. De même, lorsque la communication porte sur un thème particulier (politique, sport…), la probabilité est forte que les prochains mots de contenu relèveront du même champ sémantique que les mots précédents. Une adaptation dynamique de la prédiction au thème courant du discours pourrait dès lors apporter une aide appréciable.

Une expérience nous a montré l'extrême influence de ces différents facteurs (Wandmacher, Antoine 2006). Nous avons testé la version française de Sibylle entraînée sur le journal *Le Monde* sur des corpus correspondant à des genres et à des auteurs différents. Les résultats (tableau 2) montrent que la dégradation des performances peut atteindre presque 17 points de

KSR. Aussi est-il nécessaire de mette en œuvre une prédiction sachant s'adapter au contexte de communication courant. C'est ce que cherchons à faire dans notre système.

Corpus de test	Journal	article scientifique	roman	dialogue oral	courriel
KSR	50,5%	33,9%	40,3%	35,5 %	42,1%
Dégradation KSR	-	-16,6%	-10,2%	-15%	-8,4%

Tableau 2 : Performances du système Sibylle (version 2.4 : trigram de mots sans adaptation, apprentissage sur *Le Monde*) sur plusieurs corpus de test de registres et auteurs différents

2.3 SIBYMOT : adaptation à l'utilisateur

Adaptation à l'utilisateur – Selon une technique assez classique, nous avons combiné le modèle initial avec un modèle 3-gramme appris dynamiquement sur les saisies de l'utilisateur. Le premier modèle sert ainsi de référence de langue générale. Le second modèle sert de son côté d'adaptation au vocabulaire et aux tournures les plus fréquentes propres à l'utilisateur. L'influence des modèles est pondérée par interpolation linéaire, estimée par algorithme EM :

$$P(w_i|w_{i-1}...w_{i-3}) = \lambda_1 \cdot P_{général}(w_i|w_{i-1}...w_{i-3}) + \lambda_2 \cdot P_{utilisateur}(w_i|w_{i-1}...w_{i-3}))$$

Adaptation au registre de communication – Nos expérimentations ont montré que ce modèle interpolé permettait également une certaine adaptation au registre de communication. Du fait de la base d'entraînement utilisée, le modèle de langue général est spécifique au registre journalistique. Développer un modèle utilisateur sur des textes variés permet de couvrir une plus grande variété de registres. En reproduisant l'expérience du tableau 2 avec le modèle interpolé (tableau 3), on observe une réduction significative des baisses de performances sur d'autres registres. Ces performances dépassent largement celles obtenues à l'aide d'un modèle de cache (Kuhn, De Mori 1990), technique d'adaptation contextuelle fréquemment utilisée.

KSR / Corpus de test	**journal**	**article scientifique**	**Roman**	**dialogue oral (transcriptions)**	**Courriel**
N-gram sans adaptation	50,5%	33,9%	40,3%	35,5 %	42,1%
N-gram + simple cache	-	35,1%	40,8%	39,0%	43%
SIBYLLE avec adaptation	-	**43,1%**	**46,9%**	**50,1%**	**51,5%**

Tableau 3 : adaptation de Sibylle sur des corpus de registres et auteurs différents.

2.4 Adaptation au thème du discours : analyse sémantique latente

Nous nous sommes ensuite intéressés à l'adaptation au thème courant du discours. Il est évident que la probabilité d'occurrence de mots de contenu dépend fortement de ce thème. Par exemple, un mot a priori assez rare comme *contrepoint* aura une probabilité d'occurrence plus élevée dans un contexte de musique baroque. Plusieurs approches ont été proposées pour l'identification de thème, dont celui de (Bigi et al. 2001). De même, le modèle *trigger* (Rosenfeld 1996 ; Matiasek & Baroni 2003) utilise des collocations pour s'adapter implicitement au thème. Dans ce modèle un mot déclencheur augmente (dès qu'il est utilisé) la probabilité d'autres mots associés.

Les gains apportés par ces modèles restent toutefois limités. C'est pourquoi nous avons étudié un nouveau modèle d'adaptation thématique, basé sur l'analyse sémantique latente (LSA pour *Latent Semantic Analysis*). La LSA (Deerwester et al. 1990) a surtout été utilisée en recherche d'information sous le nom de *Latent Semantic Indexing* (LSI). Elle consiste à représenter la sémantique d'une collection de documents par une matrice [termes x documents]. Chaque élément de la matrice a pour valeur la fréquence normalisée (TF/IDF) d'apparition d'un terme dans un document donné. Les termes, qui jouent le rôle de mots-clefs décrivant l'espace sémantique, correspondent aux mots de contenu les plus fréquents de la langue. On travaille ainsi sur une matrice de très grande dimension qui est réduite par une décomposition en valeurs singulières. Cela permet de se limiter à un espace de 200 à 300 dimensions portant la majeure partie de l'information. Au final, la sémantique de chaque mot ou groupe de mots peut être représentée par un vecteur dans cet espace. On peut dès lors faire des estimations de proximité sémantique à partir, par exemple, du cosinus de l'angle formé par deux vecteurs de l'espace.

Nos travaux constituent la première tentative d'adaptation de la LSA à la prédiction de mots. La notion de document n'étant pas essentielle dans ce cadre applicatif, nous partons d'une matrice [mots_clés x termes] qui requiert une base d'apprentissage moins volumineuse. Nous supposons que le thème du discours, ou plutôt le champ sémantique qu'il parcourt, peut être appréhendé à partir des N derniers mots de contenus déjà saisis (N = 100). Pour la LSA, ce contexte sémantique est décrit par la somme des vecteurs des ces N mots.

La (pseudo-)probabilité d'occurrence d'un terme compte tenu de ce contexte est estimée par la proximité sémantique entre le vecteur de ce mot et le vecteur contexte. Celle-ci est estimée à partir du cosinus de l'angle que forment les deux vecteurs :

$$P_{LSA}(w_i|h) = \frac{\left(\cos(\vec{w}_i,\vec{h}) - \cos_{\min}(\vec{h})\right)^{\gamma}}{\sum_k \left(\cos(\vec{w}_k,\vec{h}) - \cos_{\min}(\vec{h})\right)^{\gamma}}$$

Le terme γ d'élévation à la puissance est un facteur de température qui sert à renforcer les contrastes de valeur. La distribution des probabilités sémantiques observées est en effet assez plate. Cette probabilité sémantique est interpolée avec celle du modèle markovien de langage. La combinaison la plus efficace repose sur l'utilisation d'une interpolation géométrique :

$$P'(w_i) = \frac{P_b(w_i)^{\lambda_1} \cdot P_s(w_i)^{(1-\lambda_1)}}{\sum_{j=1}^{n} P_b(w_j)^{\lambda_1} \cdot P_s(w_j)^{(1-\lambda_1)}}$$

Celle-ci a pour intérêt de ne favoriser que les hypothèses lexicales qui sont cohérentes d'un point de vue « syntaxique » (modèle de langue) et sémantique.

Une dernière modification a été apportée au modèle LSA. Wandmacher (2005) a montré que les relations sémantiques identifiées par la LSA étaient d'autant plus pertinentes que les mots considérés disposaient de voisins sémantiques rapprochés. Nous avons donc décidé de renforcer la probabilité des mots situés dans des zones de forte densité sémantique. Cette densité est estimée que la distance moyenne des N plus proches voisins d'un terme :

$$D_m(w_i) = \frac{1}{m} \cdot \sum_{j=1}^{m} \cos(\vec{w}_i, NN_j(\vec{w}_i))$$

En pratique, nous calculons cette mesure de confiance pour les cent plus proches voisins. Le coefficient d'interpolation entre LSA et modèle de langage n'est alors plus basé sur un

paramètre global obtenu par maximisation de vraisemblance (algorithme EM). Pour chaque mot, nous le calculons par la formule $\lambda_i = \beta \cdot D(w_i)$ avec β = 0.4 dans notre cas.

L'espace sémantique de notre application a été calculé à l'aide du toolkit *InfoMap*[1] à partir d'un extrait de 100 000 000 de mots du corpus *Le Monde*, en utilisant comme mots clés les 3000 mots lexicaux les plus fréquents dans le corpus et une réduction de l'espace à 150 dimensions après décomposition en valeurs singulières.

Corpus de test	4-gram	4-gram + cache	SIBYLLE
KSR	54,39%	54,6%	55,65%
Dégradation KSR	-	+0,21%	+1,26%

Tableau 4 : Performances du système Sibylle avec adaptations utilisateur et sémantique. Apprentissage sur le journal *Le Monde* et test sur un extrait du journal l'*Humanité*.

Le modèle de prédiction obtenu permet une augmentation très significative des performances (tableau 4). Là encore, ses capacités d'adaptation sont significativement supérieures à celle du modèle cache. Le système est réellement utilisable au quotidien, pour tout type d'application. Nous portons actuellement nos efforts surtout sur l'interface utilisateur du système.

3 Interface utilisateur : vers une AAC totalement libre d'usage

La première version de SIBYLLE (N-gram sans aucune adaptation) a été confiée à des patients du centre de rééducation de Kerpape (Schadle et al. 2004). SIBYLLE est en particulier utilisé par les enfants IMC de l'école primaire intégrée au centre. De l'avis général, l'apport de SIBYLLE par rapport aux communicateurs à défilement ligne-colonne est manifeste. La phase d'apprentissage de cette nouvelle aide est simple et rapide. Le défilement et le saut de sous-clavier n'a pas posé de problème particulier. SIBYLLE est essentiellement apprécié en termes de confort, le défilement linéaire et son unique validation sont très appréciés. Les enseignants ont également constaté que les enfants composent plus de textes et font moins de fautes.

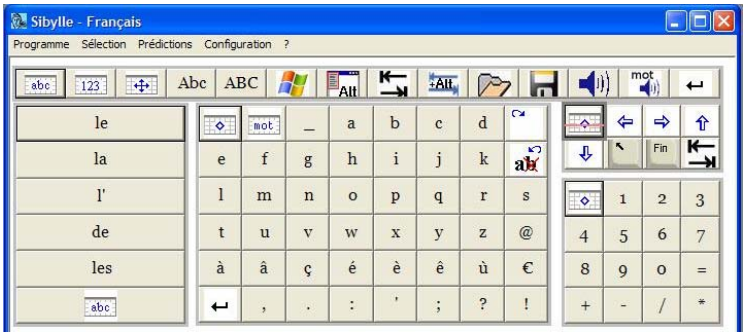


Figure 3 : Interface de la version 2.7 de SIBYLLE, utilisable avec toute application Windows

[1] http://infomap-nlp.sourceforge.net/

L'interface utilisateur utilisée (figure 2) ne permettait toutefois qu'un usage en communicateur direct : les patients saisissent leur texte pour le vocaliser. Ils utilisent également SIBYLLE pour faire des exercices sous la direction des enseignants et des orthophonistes, qui sauvegardent leur travail en fin de séance par copier-coller dans un éditeur de texte. Compte tenu des retours positifs obtenus, nous avons voulu étendre l'usage du système à toute application Windows (éditeur de texte, navigateur, de messagerie…). Le clavier a donc été étendu pour passer d'une simple saisie orthographique à un clavier complet (figure 3) interfacé avec Windows XP.

Ce passage à une version étendue nous a conduit à réfléchir aux faiblesses ergonomiques de l'interface proposée précédemment. Après discussions avec les ergothérapeutes, les enseignants et les patients, nous avons pu déterminer un ensemble de problèmes à résoudre. Ceux-ci semblent parfois relever du détail, mais l'efficacité d'une interface se niche souvent dans les détails. Ils correspondent en tous cas à des besoins essentiels des patients handicapés :

- **Paramétrisation** – Tout nouvel utilisateur handicapé pose un problème spécifique en terme d'interface, qui doit donc pouvoir être adaptée autant que possible. Aussi avons-nous étendu le nombre de fonctions paramétrables. Parmi celles-ci, on peut citer la possibilité de distinguer clic court, clic long et clic très long (durée paramétrable) et surtout d'associer à ces événements des actions spécifiques choisies parmi une liste de possibilités (effacement, mise en majuscule, retour à la ligne, basculement sur la liste de mots, vocalisation…). Lorsque la personne handicapée peut maîtriser son geste, ce mode d'entrée enrichi est très appréciable.

- **Modes de sélection** – Si le défilement linéaire est efficace avec SIBYLETTRE, certains patients, en nombre limité, ne supportent pas la réorganisation dynamique du clavier du fait de troubles associés de la vision. Nous avons donc rajouté un mode classique de sélection ligne-colonne sur clavier statique qui peut être choisi par paramétrage.

- **Défilement curseur** – Il est également apparu que la vitesse de défilement était la cause d'un problème inattendu : le saut soudain du curseur d'un symbole à l'autre engendre un stress chez l'utilisateur lorsqu'on approche de la touche recherchée. Cela conduit à des erreurs de sélection, trop rapide ou trop tardive. Pour permettre au patient de préparer son geste, nous avons ajouté une petite barre mobile qui parcourt le touche verticalement de manière cyclique : positionnée en haut de la touche après un saut de curseur, elle va descendre pour atteindre le bas juste avant le prochain saut.

- **Liste d'hypothèses lexicales** – Appris sur corpus journalistique, notre module de prédiction dispose d'un dictionnaire bien plus riche que celui des enfants ou des adolescents IMC. Afin de ne pas proposer des mots qui sont inconnus de l'utilisateur, il est possible de limiter l'affichage aux mots les plus courants de la langue. Le seuil d'affichage est paramétrable et peut donc évoluer au fil du développement cognitif.

- **Réorganisation des claviers** – Enfin, nous avons revu le positionnement de certaines touches sur le clavier afin d'arriver à une plus grande cohérence dans l'application. Par exemple, un sous-clavier ne mélangera plus des parties statiques et dynamiques, à l'exception des touches de basculement de sous-clavier.

La nouvelle version du système a été installée au centre de Kerpape en mars 2007. Nous n'avons pas encore un recul suffisant pour discuter de son appropriation par les utilisateurs. Nous rencontrons par ailleurs quelques difficultés à piloter certains aspects particuliers des applications Windows (version XP). Les modifications ergonomiques réalisées répondaient néanmoins à des besoins clairement identifiés, de même que nos techniques d'adaptation de la prédiction conduisent à des économies de saisie très appréciables. Pour autant, des améliorations sont encore envisageables.

4 Perspectives : projet ESACIMC

Trois pistes d'amélioration semblent devoir être explorées en priorité. La première concerne uniquement la prédiction de mots du point de vue de l'adaptation au registre de langage utilisé. Il est évident qu'on n'emploie pas le même style de rédaction dans une lettre officielle que dans un courrier électronique, et qu'on gagnerait à identifier automatiquement le type de communication en cours pour spécialiser la prédiction sur un modèle de langage approprié. Nous allons aborder cette question par une voie détournée : la nouvelle version du système SIBYLLE est capable de détecter la nature de l'application utilisée pour la saisie (interface de communication du système, éditeur de texte, messagerie). A l'aide de cette information, nous allons pouvoir faire un apprentissage dynamique du modèle de langage non plus par utilisateur, mais par couple (utilisateur, application). Le système SIBYLLE choisira ainsi le modèle le plus approprié à un instant donné. Celui-ci restera bien entendu interpolé avec le modèle de langue général et le modèle sémantique LSA.

Jusqu'ici, nous avons cherché à optimiser séparément le module de prédiction et l'interface utilisateur du système. Les deux problèmes étant bien entendu interdépendants, nous cherchons désormais à mieux intégrer les résultats de la prédiction dans la conception de l'interface, suivant une démarche centrée utilisateur. Nous travaillons ainsi sur les questions suivantes :

- Le clavier dynamique des caractères est réorganisé en fonction de la probabilité d'apparition d'une lettre connaissant les quatre précédentes. Mais ces probabilités sont estimées par un modèle de langue général. Ne serait-il pas plus utile d'intégrer dans ce calcul les résultats de la prédiction de mots ?
- Les mots prédits sont présentés dans un sous-clavier à part qui est accédé au moyen d'un appui supplémentaire. Ne serait-il pas possible de l'intégrer efficacement au clavier de lettres ? Suivant quelles modalités ? Est-il utile d'afficher des mots de longueur très réduite, pour lesquels l'économie de saisie sera toujours limitée ? Quelle est la taille la plus appropriée pour la liste des prédictions lexicales ?
- Les erreurs de sélection ou les fautes d'orthographes étant nombreuses, comment placer le plus judicieusement possible les touches d'effacement ou de correction ?

On ne peut répondre à ces questions sans connaître les usages réels des systèmes : des tests sur des données artificielles comme des corpus journalistiques ne seraient pas informatifs. Aussi est-il essentiel de recueillir des corpus de texte saisis par des personnes handicapées sur des communicateurs opérationnels. Ce sujet est d'autant plus important que les patients souffrent souvent de troubles langagiers associés qui rendent leur productions fortement agrammatiques. L'objectif du projet ESACIMC (www.irit.fr/ESACIMC/), soutenu par la Fondation Motrice, est précisément de recueillir des corpus réels qui seront mis en regard du tableau clinique des personnes « enregistrées ». Ce projet réunit des concepteurs de systèmes de communication assistée (laboratoires LI, VALORIA, IRIT), le centre de rééducation de

Kerpape, ainsi qu'un laboratoire (Jacques Lordat) qui s'attachera à relier tableaux cliniques et troubles langagiers.

Afin d'exploiter au mieux des données recueillies, nos systèmes ont été équipés d'un module de trace qui enregistrera toutes les actions de l'utilisateur et les réactions du système (organisation du clavier dynamique, liste de prédiction de mots…). Cette trace permettra un suivi très fin du comportement du système. Elle nous permettra de mieux caractériser les usages réels des systèmes, mais également de rejouer les interactions tracées (simulation) pour étudier l'influence de telle ou telle modification des systèmes. Ces fichiers de log respectent un format XML commun à tous les participants, ce qui favorisera leur réutilisation. Dans un avenir plus lointain, on pourrait même imaginer que l'analyse en continu de ces traces permettrait au système de s'adapter automatiquement au comportement de l'utilisateur.

Remerciements

Ce projet est partiellement financé par la Fondation Motrice (projet ESACIMC). Tous nos remerciements à Jean-Paul Departe et ses collègues du centre de Kerpape.

Références

BERARD C. (2004). Clavier-écran: concevoir avec les utilisateurs. *Handicap'2004*, Paris, 83-88

BIBER D. (1993). Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics*, 19(2), 219-241

CANTEGRIT B., TOULOTTE J.-M. (2001). Réflexions sur l'aide à la communication des personnes présentant un handicap moteur de la communication. *TALN 2001*. vol. 2, 193-202.

DEERWESTER S. C., DUMAIS S., LANDAUER T., FURNAS G., HARSHMAN R. (1990). Indexing by Latent Semantic Analysis, *JASIS* . 41(6), 391-407

GOODMAN J. (2001). A Bit of Progress in Language Modeling, *Microsoft Research Technical Report MSR-TR-2001-72.*

KUHN R., DE MORI R. (1990). A Cache-Based Natural Language Model for Speech Reproduction, *IEEE Trans. on Pattern Analysis and Machine Intelligence*.12 (6), 570-583

MATIASEK J., BARONI M. (2003). Exploiting long distance collocational relations in predictive typing. *EACL-03 Workshop on Language Modeling for Text Entry Methods*, Budapest. 1-8.

ROSENFELD R. (1996). A maximum entropy approach to adaptive statistical language modeling. Computer *Speech and Language*. 10 (1), 187-228.

SCHADLE I., ANTOINE J.-Y., LE PÉVÉDIC B., POIRIER F. (2004). Sibyl - AAC system using NLP techniques. Actes *ICCHP'2004*, Paris, France. In *LNCS 3118*, Springer Verlag.

STOLCKE A. (1998). Entropy-based pruning of backoff language models. Actes *DARPA Broadcast News Transcription and Understanding Workshop*. 270-274.

STOLCKE, A. (2002). SRILM - An Extensible Language Modeling Toolkit. Actes *ISCLP 2002, Intl. Conf. Spoken Language Processing*, Denver, Colorado.

VELLA F., VIGOUROUX N. (2007). Layout keyboard and motor fatigue: first experimental results. AMSE-journals (Barcelona, Espagne), Modelling C, Vol. 67. April 2007. 22-31

WANDMACHER T. (2005). How semantic is Latent Semantic Analysis? Actes *TALN/RECITAL 2005*, Dourdan, France.

WANDMACHER T., ANTOINE J.-Y. (2006). Training language models without appropriate language resources: experiments with an AAC system for disabled people. Actes *LREC'2006*, Genova, Italie.

De l'*amorçage* d'idées à la *composition* et *expression* de messages

Michael ZOCK
Laboratoire d'Informatique Fondamentale (LIF), CNRS, UMR 6166,
Case 901 - 163 Avenue de Luminy, 13288 Marseille
Michael.Zock@lif.univ-mrs.fr

Résumé. L'objectif de cette communication est d'explorer un problème rarement abordé, pourtant capital, en génération du langage : l'*entrée conceptuelle*. Comment aider le rédacteur à *concevoir* et *composer* un message, afin que l'ordinateur puisse le traduire en langue ? Il y a en particulier deux problèmes à résoudre, celui d'*accès* aux mots-concepts et celui de leurs *combinaisons* en messages (construction). Nous présenterons ici deux méthodes pour accéder aux éléments à partir desquels sera construit ce message : l'amorçage associatif, raccourci utilisé pour accéder aux mots-concepts (termes recherchés), et navigation dans une ontologie linguistique construite à cet effet. Etant donné que cette ressource est en cours de construction, nous présenterons ici seulement les grands principes de sa construction et quelques problèmes que pourrait poser son usage.

Abstract. The goal of this paper is to deal with a problem hardly ever addressed in the literature on natural language generation, despite its importance: conceptual input. How can we help an author to *conceive* and *compose* a message so that the computer can translate it into a natural language? There are two problems here at stake: access of words and/or concepts, and their *combination* into a message. We will present here two methods for accessing the elements from which the message is to be built: associative priming, i.e. navigation in a huge associative network, a shortcut used in order to speed up finding the token(s) the author is looking for (concept-word), and navigation in a linguistically motivated ontology, built for this task. Given the fact that the latter resource does not exist yet, we are just about to build it, we can present here only the underlying principles of building and using it and highlight some of the problems that might arise in doing so.

Mots-clés : accès/amorçage/émergence/conception/gestation d'idées, remue meninge, composition de message, navigation, idées, espace conceptuel.

Keywords: ideation, brainstorming, conceptual bootstrapping, message composition, navigation, conceptual space.

1 Introduction : problème et objectif

Produire du langage consiste essentiellement à concevoir, traduire (formuler) et exprimer oralement ou par écrit une, voire plusieurs idées ou messages.[1] Il y a donc différentes tâches ou niveaux : le conceptuel (fond), le linguistique (forme) et le physique (signal). Nous nous intéressons ici essentiellement au premier, celui des mots-concepts ou concepts-mots et leur combinaison en messages[2], la question étant, quelles connaissances donner à la machine afin qu'un utilisateur puisse lui communiquer librement ses idées afin qu'elle puisse les traduire en langue ? L'accent est donc mis sur les aspects conceptuels (conception) et non pas sur les aspects linguistiques (formulation et expression/synthèse).

Autrement dit, nous aimerions assister un être humain à composer le message que la machine est sensée traduire en Français. Idéalement, l'utilisateur devrait pouvoir communiquer librement (sans contraintes d'ordre) n'importe quel message, de n'importe quelle complexité, en clair, toutes les idées (messages) concevables dans une langue. A cette fin nous construisons une interface (ontologie linguistique) grâce à laquelle l'utilisateur peut composer progressivement son message. Cela suppose la résolution d'au moins quatre types de problèmes :

- détermination des concepts, éléments à partir desquels on construit le message ;
- règles de bonne formation : un message doit être sensé, complet et bien formé ;
- organisation des concepts afin de faciliter l'accès et la navigation ;
- représentation du message (ici, un réseau sémantique ou graphe conceptuel).

Le problème est loin d'être trivial, surtout si l'on souhaite que même un utilisateur non spécialement formé à cette tâche puisse communiquer librement (donc, dans n'importe quel ordre) toutes ses idées, quelle que soit leur complexité. Le problème est encore plus complexe si l'on veut offrir toutes ces possibilités à quelqu'un ayant certaines déficiences mentales ou linguistiques. Par ailleurs, la difficulté croît en fonction des possibilités offertes en termes de (a) la *couverture* conceptuelle (quelques idées ou toutes les idées), (b) la *complexité* d'idées (phrases simples ou complexes) et (c) la *flexibilité* offerte (présence ou absence de restrictions d'ordre) pour communiquer à la machine des fragments d'idées.

Nous procéderons ici en sens inverse de l'ordre naturel de la production : commençant par l'*expression*, nous discuterons ensuite des problèmes liés à la *composition* et l'*idéation* (conception) de messages. Aussi nous commencerons par une brève présentation et analyse du logiciel Illico, pour montrer ensuite comment l'enrichir en structurant conceptuellement

---

[1] Les *messages* sont l'expression des *idées*, qui, elles, correspondent à ce que les psychologues appellent 'proposition', c'est-à-dire, une structure dont la forme la plus élémentaire est un prédicat et un argument, par exemple : `ronde (terre)`, qui pourrait s'exprimer en Français par '`la terre est ronde`'. C'est donc avec beaucoup de reticence que nous utiliserions le terme d'idée pour faire référence aux mots ou leurs équivalents conceptuels. Quant aux termes 'concept-mot' ou 'mot-concept' nous entendons par là un concept pour lequel il y a un mot en langue, c'est un concept lexicalisé. Par exemple, le mot-concept « cosy » de l'anglais n'a pas de correspondant en français, alors qu'il en a en allemand : gemütlich.

[2] Bien que tous les systèmes de génération partent du postulat que les idées précèdent leur expression, les aspects conceptuels, à savoir, la définition, l'accès, la réprésentation ou la construction d'idées, sont largement négligés. Ceci est aussi bien vrai en psycholinguistique (Levelt, 1989) qu'en linguistique informatique (Bateman et Zock, 2003 ; Reiter et Dale, 2000).

les mots parmi lesquels l'utilisateur doit choisir afin de continuer le message produit. Puis nous terminerons par l'exposé des principes sousjacents à la future interface conceptuelle et de la méthode employée pour aider le rédacteur à accéder aux termes à partir desquels il va composer son message.

2 Un point de départ : Illico

ILLICO[3] est une plate-forme logicielle permettant de développer des applications dans le domaine de l'analyse et de la synthèse du langage naturel. C'est surtout la synthèse qui nous intéresse ici, plus précisément, la complétion de message. La copie d'écran (figure-1) montre ILLICO en action. Elle est composée de trois cadres (frames, fenêtres), signalant respectivement les mots-candidats parmi lesquels choisir pour continuer la chaîne commencée (colonne de gauche : café, chat, chien, chocolat...), diverses représentations des niveaux intermédiaires (grande fenêtre : ontologie, réprésentation sémantique et syntaxique),[4] et la forme produite jusqu'à ce point là (ligne en bas de la figure-1 : *Lea photographie le* ... < ? >).

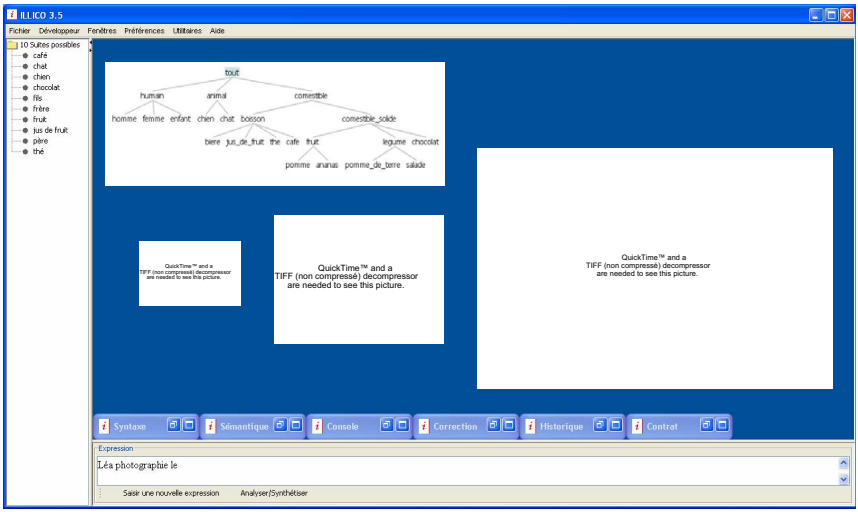


Figure-1 Illico en action : "*Lea photographie le...* < ? > »

La composition guidée par synthèse partielle est un des modes d'utilisation d'ILLICO. Il repose essentiellement sur le principe de la complétion. Le système donne la priorité aux aspects séquentiels (le mot suivant) plutôt qu'aux aspects hiérarchiques (dépendance), proposant alors le *déterminant* (le, la, les) avant le *nom* (tête) dont il dépend. Pour une discussion plus approfondie, voir (Zock, Sabatier et Jakubiec-Jamet, 2007). La question est

[3] Pour plus de détails sur ILLICO comme outil d'aide aux handicapés ou comme système en général (aspects techniques, philosophie sous-jacente), voir respectivement (Guenthner et al. 1992 ; Pasero et al. 1994) et http://www.lif-sud.univ-mrs.fr/~paulsab/ILLICO/illico.html

[4] L'arbre dérivationnel (figure à droite) ne correspond pas tout à fait à la phrase en cours.

donc de savoir comment passer d'un système de complétion d'expressions à celui d'un système d'aide à la composition de messages ? Le problème de la complétion d'une phrase est un peu différent de celui de la construction du message sous-jacent. Il s'apparente davantage à celui d'expression qu'à la construction du message. Or, avant de pouvoir exprimer quoi que ce soit en langue, il faut déjà avoir quelque chose à dire. C'est un problème de conception. Dans ce contexte, trois cas de figure nous intéressent en particulier :

1. On sait à peu près ce qu'on veut dire, mais ne l'ayant pas extériorisé, l'interprête, en l'occurrence l'ordinateur, ne pourra le traduire en langue. Le but est donc d'aider l'être humain à extérioriser sa pensée.
2. On connaît assez bien les fragments conceptuels (saluer, Yves, Jean), mais on n'a pas encore décidé comment les relier. C'est donc un problème de composition : bonne formation, adéquation par rapport à l'intention.
3. Ne sachant pas en détail tout ce qu'on aimerait dire, mais connaissant les grands lignes et disposant de quelques fragments, on aimerait que l'ordinateur nous aide à préciser notre pensée et à trouver les éléments manquants.

Il y a donc trois objectifs différents : aider à extérioriser une pensée, assurer sa bonne formation, et aider à compléter et préciser une pensée.

3 Composition du message en naviguant dans une ontologie linguistique

Comme déjà mentionné, afin de permettre à une machine de traduire nos idées en langue, il faut déjà les lui avoir communiquées. C'est à cette fin que nous construisons une ressource permettant à un être humain de composer son *message*. Plus précisément, nous aimerions concevoir une interface conviviale et dynamique dont la sélection des différents éléments du message soit guidée par un arbre dont les *feuilles* sont des concepts/mots (*vin*, *fromage*, *vélo*, *voiture*)[5] et les *noeuds* les noms des classes conceptuelles correspondantes, généralement des hyperonymes (*nourriture*, *moyens de transports*). C'est en naviguant dans cette arborescence qu'on communiquera progressivement les différents éléments (entrée), c'est-à-dire des mots/concepts, qui, une fois connectés (ou inserrés au bon endroit) forment le message. La sortie, le message, se fera incrémentalement, et sera constituée d'un graphe mettant en relation les concepts/mots. Les nœuds du graphe-message correspondent généralement à un parcours complet dans l'arborescence.

La ressource en question est une ontologie linguistique,[6] couplée à un dictionnaire et une interface graphique pour aider les entrées (options conceptuelles) et sorties (résultats des choix : graphe message). Contrairement aux autres ontologies, celle-ci a pour but d'aider la construction de message, ce qui impose des contraintes particulières, notamment le couplage de l'ontologie à un dictionnaire, contenant des informations concernant des termes (traits

[5] Que les feuilles soient des mots, des grappes de synonymes ou des concepts-mots est important à la fois d'un point de vue théorique et pratique. Nous avons opté ici pour des *synsets* à la manière de WordNet (Miller, 1990). Hélas, faute de place nous ne pouvons justifier ici notre choix, cependant nous discuterons une des raisons de cette approche dans la section 3.2 (voir le 3ème paragraphe).

[6] Nous voulons dire par là, qu'à chaque différence de sens correspond un nœud dans l'arborescence. Autrement dit, il n'y a qu'un seul nœud pour des synonymes : celui de sa classe.

sémantiques) et de leur combinabilité (valence, rôles sémantiques). C'est grâce à ces informations et les choix effectués par l'utilisateur que le système parviendra à construire le graphe message.

Le logiciel développé devra faire apparaître deux espaces : l'un pour représenter l'arborescence, dont le but est de guider l'utilisateur à trouver les briques (concepts/mots) avec lesquels le système est sensé construire le message (entrée), et l'autre, le message construit incrémentalement (graphe), en fonction des choix de l'utilisateur (sortie). La figure 2 illustre comment, partant des choix de l'utilisateur, le système construit en trois étapes le contenu du message suivant : (1) Zidane (2) a catapulté la balle (3) droit dans le filet.[7]

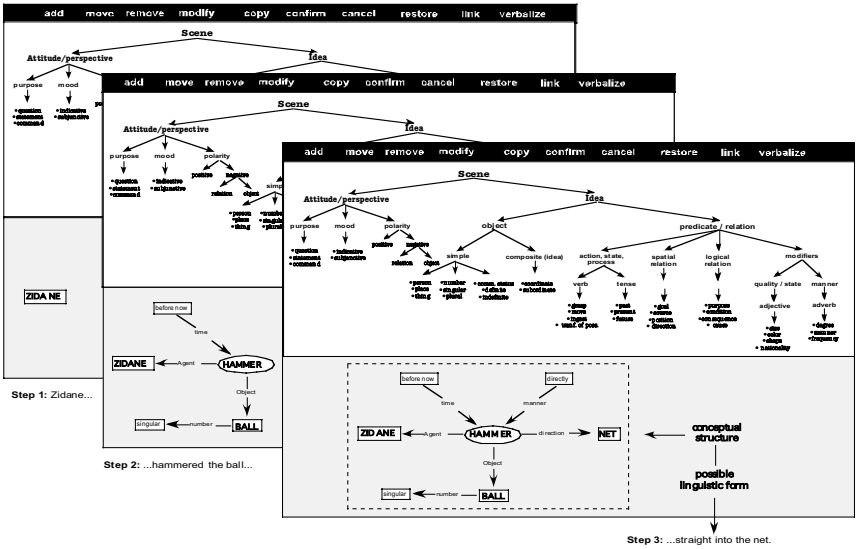


Figure 2 : construction incrémentale du contenu du message :
(1) Zidane (2) a catapulté la balle (3) droit dans le filet.

Si l'objectif de cette interface est relativement claire, on peut se demander selon quels principes elle est construite et quels sont ses avantages ou inconvénients en termes d'usage.

3.1 Principes de construction

Partant d'un corpus de mots (approche sémasiologique, bottom-up)[8] on caractérise ces derniers en termes d'inclusion. On forme donc des classes auxquelles on donne un nom.

[7] Cette interface s'inspire de Swim (Zock, 1993), générateur interactif destiné à des apprenants d'une langue étrangère.

[8] Nous partons ici des mots d'une méthode de langue et ceux d'un domaine restreint, le football (Sabatier, 1997), pour indexer ensuite ceux d'une encyclopédie destinée à des enfants. Des encyclopédies destinées aux adultes, ou des domaines ouverts seront donc exclues dans un premier temps.

Comme ces classes peuvent être groupées à leur tour, on arrive progressivement à une hiérarchie de types (pomme-fruit-objet comestible...) dont la racine domine le tout. C'est un arbre à la manière de ceux de Porphyre, philosophe grec vivant de 232 à 304. A noter cependant, que dans notre cas la racine ne s'appelle pas 'tout' mais 'scène'. En effet, partant du postulat, que l'unité de communication n'est pas le 'tout', mais une 'scène', catégorie dominant des éléments appelés *idée* et *point de vue*, nous avons ici des catégories bien différentes de celles qu'on trouve normalement dans une ontologie. Les niveaux supérieurs entretiennent donc des relations autres que celles d'un arbre à la Porphyre. Au lieu d'être du type inclusif, ils sont plutôt du type méronymique (partie-de).

Nous considérons donc, qu'un locuteur conceptualise le contenu de son message en termes de scène, qui est composée d'une *idée* (l'essentiel du message) et d'un *point de vue* (en gros, l'acte de parole). Quant à l'*idée*, elle se décompose en un *objet* (en langue, correspondant souvent à la catégorie syntaxique du nom : enfant) et en un *predicat* (par example : petit) ou d'une *relation*. Nous avons groupé ici des catégories ontologiques aussi différentes que celles correspondantes en langue aux verbes, adjectifs, adverbes, prepositions, et aux connecteurs. La raison pour cela est simple : ils ont tous plus ou moins la même fonction, à savoir, qualifier un terme (nom ou verbe dans le cas des adjectifs et adverbes) ou relier plusieurs termes (verbes, préposition et connecteur, reliant respectivement des noms ou des phrases). Ce n'est donc qu'à partir de ce niveau qu'on retrouve en gros ce qu'on va trouver dans une ontologie linguistique classique.[9]

3.2 Principes d'usage

Nous avions évoqué la possibilité de permettre une navigation libre, permettant de commencer par n'importe quelle catégorie ontologique, par exemple, un objet, une action ou un attribut (typiquement exprimés par un nom, un verbe, ou un adjectif). C'est d'ailleurs ce qui semble se passer en discours spontané, la précédence prenant le pas sur la dominance. En effet, en produisant quelque chose comme « le petit chat miaule » on commence par l'article, pour articuler ensuite l'adjectif, puis le nom, etc. Ceci dit, c'est trompeur et cela présente certains inconvénients, dans la mesure où la forme des déterminants et des adjectifs dépend de certaines caractéristiques du nom (notamment celle du genre et nombre), qui lui doit donc être déterminé avant.

Il y a une autre raison de justifier cette stratégie. Pour éviter de noyer l'utilisateur sous une masse d'informations peu pertinentes à un moment donné (par exemple, si on lui présente toute l'ontologie à la fois), il est préférable de lui présenter uniquement les choix utiles à ce moment. Autrement dit, des informations relatives aux adjectifs ou déterminants ne seront pertinentes qu'à partir du moment où l'utilisateur a décidé qu'il veut parler d'un objet (donc, d'une entité susceptible d'être exprimée sous forme d'un nom) plutôt que d'une action. On présentera donc l'objet comme *noyau* et les informations afférentes (déterminants, adjectifs, etc.) comme *satellites*.

Enfin, il y a un autre aspect intéressant concernant les stratégies de navigation. On pourrait concevoir les n nœuds de l'arbre et les concepts-mots (feuilles) comme deux espaces ou vases

---

[9] Nous utilisons ici le terme *linguistique* pour signaler qu'on part des faits de la langue : on classe ou catégorise des lemmes ou des mots, qui, eux, constituent la matière première et notre point de départ.

communiquants, ce qui permet de s'arrêter en principe à n'importe quel niveau, donc, avant d'être arrivé au niveau le plus bas, celui des feuilles. Ceci peut être vu comme un avantage ou un inconvénient. Voyons donc les conséquences pour le choix des concepts-mots et la construction du message. Le manque de précision conceptuelle (sous-détermination) aura pour effet qu'il faudra choisir ensuite dans une liste plus ou moins large. Une grande liste peut être vue comme un inconvénient (un peu ce que nous avions ressenti en regardant la colonne gauche de la figure 1, dans Illico), mais cela a aussi un avantage : on peut continuer le traitement (construction du message) sans être bloqué. Certes, à ce point précis le message manque de précision, mais c'est là justement un des avantages. Le système notera l'étiquette qu'on a su déterminer à ce moment (par exemple : *fruit*), permettant d'y revenir plus tard, pour préciser alors qu'il s'agissait d'une 'pomme'. Autrement dit, on peut combiner des approches de traitement en profondeur et largeur.

En conclusion, si la démarche esquissée ci-dessus nous semble adéquate pour le problème de la composition de message, elle est longue et le métalangage n'est pas forcément à la portée de tout le monde. Ces deux inconvénients risquent d'accroître avec l'augmentation de la taille de la couverture conceptuelle. J'ai d'ailleurs fait plusieurs propositions, indiquant comment éviter certains écueils liés à la technicité du métalangage et au limites des approches symboliques tout court. Elles concernaient trois catégories ontologiques : le *temps* (Ligozat et Zock, 1992), l'*espace* (Briffault et Zock, 1993) et les *questions* (Zock et Mitkov, 1991). Quoi qu'il en soit, la navigation dans une ontologie n'est d'aucun secours pour aider à trouver des co-occurrences (la fameuse madeleine de Proust). Aussi, pour accéder rapidement aux concepts, il vaut mieux utiliser une autre technique.

4 L'amorçage de concepts basé sur la notion d'associations

Il nous arrive parfois de ne pas trouver un mot ou un concept, alors que d'autres, associés plus ou moins fortement, surgissent sans y avoir été invités. Ainsi, en cherchant le mot exprimant l'idée d'une boisson forte au gout de café ('moka') les mots suivants peuvent se présenter à l'esprit : café, dessert sucré, district de Maurice, ville portuaire du Yémen, Elvire Murail ou Antoine Virnal (respectivement auteur et musicien dont le pseudonyme est Moka), etc. Ce qui vaut pour les mots vaut également pour les concepts ou les idées, le mécanisme d'accès étant le même. L'association est un mécanisme extrêmement puissant. Il est pratiquement impossible de s'en soustraire. Qu'on le veuille ou non, les choses (mots, idées, objets) n'existent pratiquement jamais seules, hors contexte. Elles jouent généralement un rôle dans une scène, participant ensemble à des histoires, et c'est aussi pour cela qu'elles s'évoquent mutuellement. Chaque chose nous fait penser à autre(s) chose(s). Il y a donc une idée ou *mot source* (m_s), celui dont on se souvient et qu'on est capable de produire, et une, voire plusieurs idées ou *mots cibles* (m_c) : le concept ou mot recherché. Entre les deux il y a une force d'attraction (force associative), variable en fonction d'un certain nombre de paramètres (notamment, le contexte).

Les mots ou idées peuvent avoir différents types de rapports : rapport de *sens* (jaune-banane; fruit-banane, chien-dog, etc.) [10]; (b) rapports de *forme* (vin vs. vingt); (c) ou les deux à la fois

[10] Il y a toute une serie de liens fréquemment utilisés, dont certains sont désormais intégrés dans des dictionnaires électroniques (1) équivalence/synonyme : aubergine-pervenche ; (2) contraire/antonyme : grand-petit ; (3) spé-cificité, hyponyme/hyperonyme : pain-nourriture ; (4) partie de/méronyme : moteur-voiture ; (5)

(chat-rat). Enfin, la distance entre les deux éléments source et cible peut-être plus ou moins grande. Aussi, au lieu de pouvoir atteindre directement la cible on peut être obligé de passer par un, voire plusieurs concepts/mots intermédiaires. Autrement dit, on doit naviguer, faisant éventuellement même des détours, l'essentiel étant d'y arriver.

Si en principe tous ces rapports peuvent nous être utiles, nos efforts actuels portent essentiellement sur l'accès par le sens, plus précisément sur les associations syntagmatiques. L'hypothèse étant, que nos connaissances (mémoire encyclopédique, mots, concepts) sont un vaste réseau dont les mots ou *concepts* sont les nœuds et les liens des *associations*. L'accès aux éléments recherchés s'effectuera alors par navigation. Par exemple, pour chercher un mot, on entre dans le réseau en donnant un mot (mot-source : m_s) dont nous pensons qu'il est assez proche du mot recherché (mot-cible : m_c) pour recevoir en sortie tous les mots associés. Si la liste contient le mot recherché la recherche s'arrête, sinon on choisit parmi ces éléments un candidat, espérant qu'il produira au tour suivant le mot convoité, car le nouveau candidat suscitera à son tour une liste de réponses (mots). Ainsi faisant on s'approche donc progressivement du mot cible, à moins d'avoir abandonné avant.

Prenons un exemple. Supposons qu'on cherche le mot ***moka*** (m_c), alors que le seul mot qui nous vienne à l'esprit (m_s) soit ***informatique***. Le système prendra alors celui-ci comme point de départ (noyau) pour présenter tous les mots (satellites) ayant un rapport direct avec lui, par exemple, *Java, Lisp* ou *Prolog* (*langages* de programmation), ordinateur, souris, imprimante (*matériel informatique*) ou Mac, PC (*types d'ordinateur*), etc. C'est à l'utilisateur de décider dans quelle direction continuer la recherche, car il n'y a que lui qui sait quel mot parmi ceux présentés par le système correspond le mieux à son idée. Partant du m_s 'informatique', il va donc se diriger vers *langages*, sachant que ce qu'il cherche est ni un *matériel informatique* ni un *type de machine*, pour trouver *Java*, qui est non seulement un *langage de programmation* mais aussi une *île*. Prenant alors *Java* comme nouveau point de départ il va obtenir *café* (puisqu'on cultive du café à Java), puis enfin *moka*, qui est un type de café. A partir du mot *Java* il aurait également pu obtenir *Kawa* (*café* en argot, et ayant une certaine ressemblance phonétique avec le mot-source) ou *Kawa Igen* (volcan sur l'île de Java). On constate, tous les chemins conduisent à Rome, ou presque. D'autre part, on note qu'au prix de très peu d'opérations on a réussi à couvrir une assez grande distance conceptuelle, tout en réussissant à trouver assez rapidement la solution. En tout cas plus rapidement que si on avait navigué dans une ontologie, car, contrairement à cette dernière, on n'est pas contraint par des hiérarchies : deux idées, de nature très différentes (cheval, musique), donc très éloignées dans une taxinomie ou hiérarchie typée, peuvent s'évoquer naturellement et quasi instantanément, du simple fait qu'elles sont liées dans le monde réel, les deux participant dans une même scène ou petit drame.

Ces types de rapport sont essentiellement syntagmatiques. Ils sont bien différents de la plupart de ceux rencontrés dans des ontologies ou une ressource du type WordNet, qui sont essentiellement paradigmatiques. Ceci dit, même FRAMENET (Johnson et al., 2001), qui pourtant contient beaucoup de liens syntagmatiques ne permet pas ce type de navigation, puisque FrameNet est avant tout une ressource linguistique et non pas une ressource encyclopédique. Or, c'est précisément ce dont nous avons besoin : un index en termes d'associations, signalant les rapports entre les objets (lemmes) du monde réel. Les gens

---

ressem-blance/analogie : chien-loup ; (6) type de/variété de : pommes-Granny ; (7) fonction/rôle joué : couteau-couper ; (8) co-occurence/locution : moteur-puissant.

ordinaires ne structurent pas les objets comme des scientifiques (par exemple, des botanistes), mais plutôt en fonction des situations dans lesquelles ils les ont rencontrés.

Bien entendu, pour permettre ce type de navigation par associations il faut d'abord créer la ressource, c'est-à-dire, prendre un corpus, extraire les collocations (ou associations), les pondérer en termes statistiques et typer les liens repérés. Idéalement, le poids (force associative) devait même changer dynamiquement en fonction du thème, car le même mot (piano) évoque des idées bien différentes selon le thème (concert, déménagement). Nous avons commencé ce travail (Ferret et Zock, 2006) en utilisant un extracteur de collocation d'abord sur un corpus du journal *Le Monde*, puis sur un autre, à notre avis plus représentatif des connaissances du monde, Wikipédia (http://fr.wikipedia.org/wiki/Accueil).

Le choix du corpus est donc très important dans la mesure où il est supposé représenter les connaissances du monde du citoyen moyen. Il doit contenir des informations diverses du type politique, historique, géographique, scientifique (lois naturelles, chimie et physique, bref, des connaissances du type encyclopédique), etc. ainsi que des connaissances liées au présent (le sports, faits divers, etc.). Enfin, si l'intuition selon laquelle le dictionnaire mental serait un vaste réseau, dont les *nœuds* sont des mots (et/ou des concepts) et les *liens* essentiellement des associations, ne date pas d'hier, il n'y a à notre connaissance aucun inventaire exhaustif (ou de classification) de ces liens. Or, connaître leur nature est l'une des conditions préliminaires pour indexer un tel dictionnaire, et c'est là un de nos objectifs des années à venir. Pour une feuille de route, voir (Zock et Bilac, 2004).

5 Conclusions

Nous nous sommes posé la question concernant les connaissances qu'un système devait avoir pour aider un individu moyen (adulte, adolescent ou étudiant de langue) à communiquer ses idées à une machine. A cet effet nous avons proposé deux solutions (a) une ontologie linguistique couplée à un dictionnaire et un générateur de graphe et (b) un index basé sur les notions d'association et de poids (fréquence). Les deux ressources sont en cours d'élaboration, et beaucoup de questions se posent à ce stade, notamment en ce qui concerne notre ontologie. Ces questions portent sur la *complétude*, (absence de certaines catégories, par exemple, celle des « circonstants »), l'*adéquation* d'*emplacement* et d'organisation des nœuds (la place de la négation est à revoir), l'adéquation des termes métalinguistiques choisis (compréhensibilité), etc. Il est clair, même si les premiers résultats paraissent prometteurs, beaucoup de travail reste à faire. Enfin, malgré certaines ressemblances aux thésaurus (Roget, 1852), aux dictionnaires analogiques (Boissière, 1862) à des des ressources électroniques comme WordNet, FrameNet et HowNet (Dong Zhendong, 2000), ou aux ontologies tout court, concernant l'accès aux mots et concepts, nous pensons que notre proposition va beaucoup plus loin, ne serait-ce que parce qu'elle permet en plus la composition de messages. On peut penser que notre approche, ne sera pas d'un grand concours pour des personnes handicapées, du moins pas la partie symbolique (l'ontologie). En revanche, les propositions faites dans (Ligozat et Zock 1992, Briffault et Zock, 1993) donnent des pistes comment éviter l'écueil des métalangages en proposant la construction de scènes basées sur la notion d'icônes et de micromondes : tout changement de la scène se traduisant par un changement de la forme au niveau de la langue, pourvu que le changement conceptuel soit pertinent. C'était justement en observant ces corrélations forme-fond que l'utilisateur était sensé apprendre les règles de la langue.

Références

BATEMAN, J. & ZOCK, M. (2003). Natural language generation. In R. Mitkov (Ed.), Handbook of computational linguistics. London: Oxford University Press pp. 284-304

BOISSIÈRE, P. (1862) Dictionnaire analogique de la langue française : répertoire complet des mots par les idées et des idées par les mots, Paris

BRIFFAULT, X. & M. ZOCK (1994) What do we mean when we say to the left or to the right? How to learn about space by building and exploring a microworld. 6th International Conference on Artificial Intelligence: Methodology, Systems, Applications, Sofia, 363-371

DONG, ZHENDONG (2000) HowNet: http://www.keenage.com.

FERRET, O. & ZOCK, M. (2006). Enhancing electronic dictionaries with an index based on associations, Coling/ACL, Sidney, pp.281-288

GUENTHNER F., KRÜGER-THIELMANN K., PASERO R. & P. SABATIER Communication Aids For ALS Patients, Proceedings of the 3rd International Conference on Computers for Handicapped Persons (ICCHP 92, Vienne), pp 303-307, 1992

JOHNSON, R, C. FILLMORE, E. WOOD, J. RUPPENHOFER, M. URBAN, M. PETRUCK, C. BAKER (2001) The FrameNet Project: Tools for Lexicon Building, http://www.icsi.berkeley.edu/~framenet/

LEVELT, W. (1989). Speaking : from intention to articulation. Cambridge, MA: MIT Press.

LIGOZAT, G. & M. ZOCK (1992) How to visualize time, tense and aspect. COLING '92, Nantes

MILLER, G.A., ed. (1990). WordNet: An On-Line Lexical Database. *International Journal of Lexicography,* 3(4).

PASERO R. RICHARDET N. & P. SABATIER Guided Sentences Composition for Disabled People, Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP 94, Stuttgart), pp 205-206, 1994.

REITER, E., & DALE, R. (2000). Building natural language generation systems. London. Cambridge University Press.

ROGET, P. (1852) Thesaurus of English Words and Phrases, Longman, London

SABATIER, P. Un lexique-grammaire du football, Lingvisticæ Investigationes, XXI:1, pp. 163-197, J. Benjamins Publishing Compagny, Amsterdam, 1997.

ZOCK, M. , SABATIER, P & JAKUBIEC-JAMET, L. (2007) Who's Next? From Sentence Completion to Conceptually Guided Message Composition, à paraître, 4th Natural Language Processing and Cognitive Science (NLPCS), Funchal, Madeira

ZOCK, M., S. BILAC (2004). Word lookup on the basis of associations : from an idea to a roadmap. *Proc. of Coling workshop* : Enhancing and using dictionaries, Genève, pp. 29-34

ZOCK, M. & R. MITKOV (1991) How to ask a foreigner questions without knowing his language : proposal for a conceptual interface to communicate thought. In, Proceedings of the Natural Language Processing Pacific RIM Symposium, Singapore

ZOCK, M. (1991) SWIM or SINK : the Problem of Communicating thought, in : Swartz, M.& M. Yazdani (Eds.). Intelligent Tutoring Systems for Foreign Language Learning: The Bridge to International Communication. Springer Verlag, Berlin, 235-247

Index par auteurs