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

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

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 : 2-9520326-8-8

Table des matières

Préface

Les conférences TALN et RÉCITAL poursuivent leur tour de France et font cette année étape dans la ville rose. Elles y sont co-organisées à l'université de Toulouse 2-Le Mirail par l'ERSS, équipe du laboratoire CLLE (Cognition, Langues, Langage, Ergonomie, UMR 5263 — CNRS, UTM & EPHE) et par l'IRIT (Institut de Recherche en Informatique de Toulouse, UMR 5505 — CNRS, INPT, UPS, UT1 & UTM). Les deux conférences connaissent un succès toujours croissant permettant de maintenir un niveau général élevé tout en rassemblant très largement la communauté du TALN francophone.

Elles sont parfaitement représentative des travaux de cette communauté dont la forte implication se manifeste tant au niveau des soumission que des relectures. Les soumissions ont été très nombreuses en 2007 puisqu'il y a eu plus de 150 propositions de communication (126 pour TALN et 30 pour RÉCITAL) provenant de 19 pays différents. Les articles soumis à TALN ont été évalués par 99 relecteurs (membres des comités de programme et de lecture). RÉCITAL a pour sa part un comité de programme de 18 membres. Nous tenons à les remercier tous car le succès de TALN et RÉCITAL doit beaucoup à la qualité des relectures et des retours dont les propositions de communication font l'objet. Le programme de TALN 2007 comporte 39 communications orales, 30 communications affichées et 4 démonstrations. Celui de RÉCITAL 2007 se compose de 9 communications orales et 10 communications affichées.

Nous souhaitons remercier très sincèrement le comité d'organisation pour son efficacité redoutable et l'ambiance amicale dans laquelle cette manifestation a été préparée. Ces remerciements s'adressent tout particulièrement à Véronique Debats (IRIT) et le Service Communication de l'IRIT, dont la contribution au succès de la conférence a été déterminante.

Farah BENAMARA, Nabil HATHOUT, Philippe MULLER et Sylwia OZDOWSKA

TALN-2007

5 au 8 juin 2007, Toulouse, France

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

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
Segmentation

Exploiting structural meeting-specific features for topic segmentation

Maria GEORGESCUL[1], Alexander CLARK[2], Susan ARMSTRONG[1]
[1] ISSCO/TIM/ETI, University of Geneva
[2] Department of Computer Science, Royal Holloway University of London
maria.georgescul@eti.unige.ch, alexc@cs.rhul.ac.uk,
susan.armstrong@issco.unige.ch

Résumé. Dans cet article, nous traitons de la segmentation automatique des textes en épisodes thématiques non superposés et ayant une structure linéaire. Notre étude porte sur l'utilisation des traits lexicaux, acoustiques et syntaxiques et sur l'influence de ces traits sur la performance d'un système automatique de segmentation thématique. Nous appliquons notre approche, basée sur des machines à vecteurs support, à des transcriptions des dialogues multi-locuteurs.

Abstract. In this article we address the task of automatic text structuring into linear and non-overlapping thematic episodes. Our investigation reports on the use of various lexical, acoustic and syntactic features, and makes a comparison of how these features influence performance of automatic topic segmentation. Using datasets containing multi-party meeting transcriptions, we base our experiments on a proven state-of-the-art approach using support vector classification.

Mots-clés : segmentation automatique en épisodes thématiques, machines à vecteurs support, dialogues multi-locuteurs.

Keywords: automatic topic segmentation, support vector machines, multi-party dialogues.

1 Introduction

Georgescul et al. (2006b) proposed a support vector machine approach to the task of text segmentation which demonstrates improvements over state-of-the-art techniques, by modeling large scale (merely lexical) features and non-linear relations in an efficient and stable way. Their experimental results showed that word distributions in texts provide relevant information for the detection of boundaries between thematic episodes in data sets covering different domains. In this paper, we put the emphasis on tackling the topic segmentation problem in the context of recorded and transcribed multi-party dialogs. In particular, we extend the work of Georgescul et al. (2006b) by exploring potential information provided by 'surface' cues in multi-party dialogues such as syntactic knowledge, cue-phrases and acoustic cues. We investigate the pertinence of these factors individually and in combination with information provided by word distributions through the intermedium of transductive support vector machines.

In order to identify boundaries, we model the thematic segmentation task as a binary classification problem. The features considered for designing the classifier are described in Section 2. In Section 3 we highlight how the classification model is constructed by using transductive support vector learning. A comparative analysis of the support vector classifier performance by using these cues is provided in Section 4.

2 Input features

As in (Georgescul *et al.*, 2006b), we consider the thematic segmentation task as a binary classification problem, where each utterance should be classified as a topic boundary or not. As explained in Section 3, we employ a support vector machine classifier which is given as input a vectorial representation of the utterance to be classified and its context. Each dimension of the input vector indicates the value of a certain feature characterizing the utterance. For utterance characterization, Georgescul et al. (2006b) only considered features based on observations of patterns in vocabulary use. Here, in addition to these lexical features, we consider meeting-specific features as described in the following.

Note that, similar types of features examined in our study have been previously proposed for analyzing discourse structure in state-of-the-art studies like those described in (Litman & Passonneau, 1995; Hirschberg & Nakatani, 1996; Galley *et al.*, 2003). These include speaker activity, discourse markers, prosodic and syntactic features.

2.1 Speaker activity

According to (Pfau *et al.*, 2001), patterns of *speech activity* are valuable data for discourse analysis. In order to explore this claim, the first pattern we chose to investigate is speaker activity. In particular, we start with the hypothesis that in meeting data the contribution of each participant in the discussion can signal a new topic. For instance, some participants could have a preference for certain subjects of discussion.

We take into account the changes in speaker activity by measuring the number of words each participant uttered before and after each utterance candidate to a thematic boundary. This is formalized in the following manner. Let A_k^s be the number of words that the participant s said in utterance u_k. For each meeting participant s and for each i-th utterance u_i, we take into account the number of words that the participant s uttered before and during u_i in an interval of size $activityWS$, by considering the vector $\vec{fl}_i^{\,s}$ as: $\vec{fl}_i^{\,s} = \left(A^s_{i-activityWS+1}, A^s_{i-activityWS+2}, ..., A^s_i\right)$. We also store in a vector $\vec{fr}_i^{\,s}$ the number of words that the participant s uttered after u_i in an interval of size $activityWS$: $\vec{fr}_i^{\,s} = \left(A^s_{i+1}, A^s_{i+2}, ..., A^s_{i+activityWS}\right)$. We then normalize the two vectors $\vec{fl}_i^{\,s}$ and $\vec{fr}_i^{\,s}$ to form two probability distributions l_i^s and r_i^s, respectively. That is, we perform the normalization by simply dividing each element in the vector by the sum of all entries in the vector.

We measure significant changes in speaker activity by using the *information radius* between the probability distributions given by the speaker activity at the left and right side of the current (u_i)

utterance:

$$IRad(l_i^s, r_i^s) = \frac{1}{2}\left[\sum_{\substack{i \\ l_i^s \neq 0}} l_i^s log\frac{l_i^s}{m_i} + \sum_{\substack{i \\ r_i^s \neq 0}} r_i^s log\frac{r_i^s}{m_i}\right] \quad (1)$$

where $m_i = \frac{l_i^s + r_i^s}{2}$ is the average distribution of the two random variables l_i^s and r_i^s.

Finally, $IRad(l_i^s, r_i^s))$ will constitute the entry for one dimension of the vectorial representation for utterance u_i.

2.2 Discourse markers

Previous studies (Litman & Passonneau, 1995; Marcu, 2000) addressed questions regarding discourse relations and their realization by discourse markers. Here, we are interested in finding those discourse markers that indicate thematic shifts in our data. We started with the following list of discourse markers that has been synthesized from a commonly used list of discourse markers: "accordingly", "actually", "after all", "also", "although", "anyway", "back to", "basically", "but", "fine", "for example", "furthermore", "generally", "however", "like", "moreover", "nevertheless", "nor", "now", "of course", "okay", "really", "similarly", "since", "speaking of", "so", "still", "that's all", "then", "therefore", "well". For each discourse marker in this list, we automatically examine if it occurs in each utterance that is a candidate for marking a thematic boundary. That is, our SVM takes as input binary features indicating whether each discourse marker occurs in the current utterance. We retain as input features to our system only those discourse markers that occur at least once in our corpus.

2.3 Syntactic features

The use of syntax-based features is to a large extent motivated by previous work (Passonneau & Litman, 1993; Litman & Passonneau, 1995) relating discourse structure and noun phrase anaphora. Regarding the pronominal reference, we are mainly following the intuitive assumption that nouns and verbs appear more frequently at the beginning of a new topic, while pronouns appear more frequently in the middle of a thematic episode.

The syntactic features considered in our study are the distributions of different part-of-speech categories before and after a potential thematic boundary. That is, we extracted frequencies of pronouns, (proper) nouns and verbs before and after each utterance candidate to a thematic boundary. For the annotation of part-of-speech information, we used TreeTagger (Schmid, 1994).

This component was formalized as follows. Let P_i, N_i, V_i be the number of pronouns, nouns and verbs, respectively, in utterance u_i. We store in a vector $\vec{fl}_i^{\,p}$ the number of pronouns occurring in utterances situated before u_i in an interval of size $synWS$: $\vec{fl}_i^{\,p} = (P_{i-synWS+1}, P_{i-synWS+2}, ..., P_i)$. We also store in a vector $\vec{fr}_i^{\,p}$ the number of pronouns occurring in utterances situated after u_i in an interval of size $synWS$: $\vec{fr}_i^{\,p} = (P_{i+1}, P_{i+2}, ..., P_{i+synWS})$. Similarly, we store in vectors $\vec{fl}_i^{\,n}$, $\vec{fl}_i^{\,v}$ the number of nouns and verbs, respectively occurring before u_i in an interval of size $synWS$ utterances. Then, the vectors $\vec{fr}_i^{\,n}$, $\vec{fr}_i^{\,v}$ will contain the number of nouns and verbs, respectively occurring after u_i in an interval of size $synWS$ utterances.

As in Section 2.1, we normalize the resulting vectors of counts $\vec{fl}_i^{\,p}$, $\vec{fr}_i^{\,p}$, $\vec{fl}_i^{\,n}$, $\vec{fr}_i^{\,n}$, $\vec{fl}_i^{\,v}$, $\vec{fr}_i^{\,v}$ to obtain probability distributions l_i^p, r_i^p, l_i^n, r_i^n, l_i^v, r_i^v, respectively. Finally, we measure changes in the distribution of pronouns, nouns and verbs at the left and right side of the current utterance by using the information radius (see Equation 1). That is, for each utterance u_i, we measure $IRad\,(l_i^p, r_i^p)$, $IRad\,(l_i^n, r_i^n)$, $IRad\,(l_i^v, r_i^v)$, which will constitute entries for three dimensions of the vectorial representation for utterance u_i (taken as input to the SVM classifier).

2.4 Silences and overlaps

Silences and overlaps, as well as other acoustic information can also give evidence whether a major topic shift occurred. In particular, studies on discourse structure (Hirschberg & Nakatani, 1996) exploit various prosodical information such as pitch range (raised at segment-initial phrases and lower at segment-final phrases), speech rate (accelerating at segment-final phrases), amplitude and contour.

We investigated the pertinence of these features with the following formalization. Let S_i be the silence duration between utterance u_{i-1} and u_i. Let O_i be the speaker overlap duration between utterance u_{i-1} and u_i. We normalize the S_i and O_i values by speaker and the resulting values $S_i^{'}$, $O_i^{'}$ are used to compute the following quantities: $sl_i = \sum_{k=i-silenceWSL+1}^{i} \left(S_i^{'}\right)$; $sr_i = \sum_{k=i+1}^{i+silenceWSR} \left(S_i^{'}\right)$; $ol_i = \sum_{k=i-overlapWSL+1}^{i} \left(O_i^{'}\right)$; and $or_i = \sum_{k=i+1}^{i+overlapWSR} \left(O_i^{'}\right)$.

We include silences and overlaps as part of the utterance context representation by considering the sl_i, sr_i, ol_i, or_i quantities as dimensions of the vector characterizing the utterance u_i.

3 Methodology

As introduced in the previous section, we employ a vectorial representation containing lexical, acoustic and syntactic information to characterize each utterance. The topic segmentation task is thus reduced to a binary classification problem: each utterance has to be classified as marking the presence or the absence of a topic shift in the text.

In order to infer eventual dependencies between the binary class label and observations of patterns (provided by the lexical, acoustic and syntactic information), we employ a discriminative approach based on transductive support vector learning. A brief overview on inductive support vector learning for topic segmentation has been described in (Georgescul *et al.*, 2006b). In this section, we give some highlights representing the main elements in using transductive support vector learning for topic segmentation.

The support vector learner $\mathcal{L}$ is given a *training set* $S_{train} = ((\vec{u}_1, y_1), ..., (\vec{u}_n, y_n)) \subseteq (U \times Y)^n$ containing n examples drawn independently and identically distributed (i.i.d.) according to a fixed distribution $Pr(u, y) = Pr(y|u)Pr(u)$. Following the transductive setting proposed by Joachims (1999), the learner is also given an i.i.d. sample, $S_{test} = (\vec{u}_1^{\star}, \vec{u}_2^{\star}, ..., \vec{u}_k^{\star})$, containing k test examples from the same distribution as $\vec{u}_1$, $\vec{u}_2$, ... ,$\vec{u}_n$. Each training example from S_{train} consists of a high-dimensional vector $\vec{u}$ describing an utterance and the class label y. The class label y has only two possible values: +1 (corresponding to a 'thematic boundary') or -1 (corresponding to a 'non-thematic boundary'). We represent each utterance instance by a feature vector $\vec{u}$ with attributes containing 'surface' meeting-specific information (as described in Section 2)

plus the attributes given by the bag-of-words representation of word frequencies, as described in (Georgescul *et al.*, 2006b).

Given a hypothesis space $\mathcal{H}$, of functions $h : U \rightarrow \{-1, +1\}$ having the form $h(\vec{u}) = sign(\langle \vec{w}, \vec{u} \rangle + b)$, the transductive learner $\mathcal{L}_{transd}$ seeks a decision function h_{transd} from $\mathcal{H}$, using S_{train} and S_{test} so that the expected number of erroneous predictions on the test examples is minimized. Using the structural risk minimization principle (Vapnik, 1998), the smallest bound on the test error is calculated by minimizing the following cost function $\mathcal{W}^{transd}$:

$$\begin{aligned} &\mathcal{W}^{transd}(y_1^*, ...y_k^*, \vec{w}, b, \xi_1, \xi_2, ..., \xi_n, \xi_1^*, \xi_2^*, \xi_k^*) = \\ &= \frac{1}{2} \langle \vec{w}, \vec{w} \rangle + C^+ \sum_{i=0, y_i=1}^{n} \xi_i + C^- \sum_{i=0, y_i=-1}^{n} \xi_i + C^* \sum_{j=0}^{k} \xi_j^*, \end{aligned} \tag{2}$$

subject to:

$$\begin{cases} y_i \left[\langle \vec{w} \cdot \vec{u}_i \rangle + b \right] \leq 1 - \xi_i \; for \; i = 1, 2, ..., n; \\ y_j^* \left[\langle \vec{w} \cdot \vec{u}_i^* \rangle + b \right] \leq 1 - \xi_i^* \; for \; j = 1, 2, ..., k; \\ y_j^* \in \{-1, 1\} \; for \; j = 1, 2, ..., k. \\ \xi_i \geq 0 \; for \; i = 1, 2, ..., n; \\ \xi_j^* \geq 0 \; for \; j = 1, 2, ..., k; \end{cases}$$

The so-called *slack variables* ξ_i and ξ_j^* are introduced in order to be able to handle non-separable data. The *regularization parameters* C^- and C^+ are tuned as described in Section 4.1.

4 Experiments and results

4.1 Parameter estimation

We train and evaluate the effectiveness of our technique on the ICSI-MR dataset (Janin *et al.*, 2004) containing transcribed multi-party dialogs.

We divide the ICSI-MR data set into two disjoint parts: a training dataset composed of 80% of the initial data set, while the remaining 20% is held out for testing purposes. That is, the training set is used to determine the best model settings for the SVM classifier, while the test set is used to determine the final topic segmentation error rate.

We select the best model parameters, by running five-fold cross validation for SVM parameter estimation, using the Gaussian RBF kernel. During this preliminary step we estimate the performance of the SVM classifier by using the precision and recall, i.e. the precision/recall-breakeven point (Joachims, 1999). The choice of binary evaluation metrics in this step was motivated by the fact that posing the topic segmentation task as a classification problem involves a loss of the sequential nature of the data, which is an inconvenience in computing the P_k (Beeferman *et al.*, 1999) or Pr_{error} (Georgescul *et al.*, 2006a) measures.

Parameter	Interval for grid search	Best window size
activityWS	$5 \ldots 50$ step 5	35 utterances
synWS	$5 \ldots 50$ step 5	30 utterances
silenceWSL	$2 \ldots 10$ step 1	6 utterances
silenceWSR	$2 \ldots 10$ step 1	3 utterances
overlapWSL	$2 \ldots 10$ step 1	2 utterances
overlapWSR	$2 \ldots 10$ step 1	4 utterances

TAB. 1 – Grid search interval over parameters involved in data representation.

Given that the data used in our experiments contains only about 0.07% utterances marking thematic boundaries relative to the total number of utterances in the corpus, we handle the imbalance between the number of positive and negative examples for the SVM classifier by using an assymetric soft margin optimization, which charges more for false negatives than for false positives. That is, we set the regularization parameter C^+ several times larger than C^-: $C^+ = \left\lceil \frac{n}{n^+ - 1} - 1 \right\rceil \cdot C^-$, where n is the total number of training examples and n^+ is the number of positive training examples.

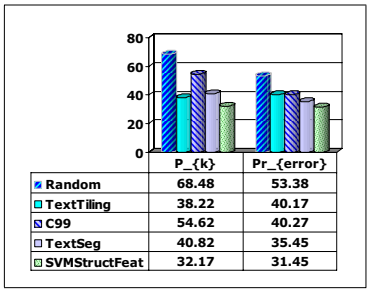
Model selection is done in two phases, as described below.

The first step in model selection consists of searching for the most appropriate utterance representation by using each individual category of features. That is, we look for appropriate values for the size of the windows (intervals) considered when measuring "speaker activity" and when taking into account "syntactic information" and "silences and overlaps" for the utterance instance (cf. Section 2). This is determined by performing a grid search interval over various values for *activityWS*, *synWS*, *silenceWSL*, *silenceWSR*, *overlapWSL* and *overlapWSR*. For each "window size (WS)" parameter, the range of values we select from is given in the second column of Table 1. Note that for the features based on lexical reiteration, we have used the optimal parameter settings that have been determined in (Georgescul *et al.*, 2006b). In this step, we train the SVMs with fixed values for both the RBF kernel parameter and the regularization parameters C^+ and C^-, i.e. the magnitude of the penalty for violating the soft margin has been set to: $C^- = 1$; while the RBF kernel parameter has been set to: $\gamma = 1$.

Using the entire set of features with the representations selected in the first step (cf. the third column of Table 1), the second step in model selection consists in optimizing the parameters of the classifier, i.e. the regularization parameters C^+ and C^- and the RBF kernel parameter γ. That is, we perform grid search interval over the following values: $C^{-1} \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, $\gamma \in \{2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 1, 2, 2^2, 2^3, 2^4, 2^5, 2^6\}$.

4.2 Results

The results obtained on the ICSI-MR corpus using only the proposed surface conversational cues, (i.e. excluding the features based on lexical reiteration), in our SVM approach for thematic segmentation are illustrated in Figure 1. The table gives means for the percentage error rates given by P_k metric (Beeferman *et al.*, 1999) and the Pr_{error} metric (Georgescul *et al.*, 2006a) for the systems we have used throughout our work. We provide as baselines the error rates obtained when using *TextTiling* (Hearst, 1997), *C99* (Choi, 2000), *TextSeg* (Utiyama & Isahara, 2001) and *Random*, a naive segmentation algorithm (by which the number of boundaries is randomly selected and boundaries are randomly distributed throughout text).



	P_{k}	Pr_{error}
Random	68.48	53.38
TextTiling	38.22	40.17
C99	54.62	40.27
TextSeg	40.82	35.45
SVMStructFeat	32.17	31.45

FIG. 1 – Comparative performance of our SVM approach using only structural features with various topic segmentation systems run on ICSI-MR data.

From Figure 1, we observe that by following the quantitative assessment of both P_k error and the Pr_{error}, our method, labeled as *SVMStructFeat*, using only surface-features outperforms other topic segmentation systems reported on in the literature.

The error values for topic segmentation on the ICSI-MR corpus when using the entire set of features (i.e. lexical, syntactic and prosodic information) are given in the first row of Table 2. The error rates of our method using both lexical and structural features, i.e the error rates of $SVM_{Lexical+StructFeat}$ in the first row of Table 2, as compared to those obtained in (Georgescul *et al.*, 2006b), i.e. the error rates of $SVM_{LexicalFeat}$ in second row of Table 2, show that performance gains can be achieved with the help of surface features in addition to word distribution-based features.

System	P_k error rate	Pr_{error} error rate
$SVM_{Lexical+StructFeat}$	20.94 %	20.17%
$SVM_{LexicalFeat}$	21.68%	21.83%

TAB. 2 – Comparative performance of our SVM approach when using only lexical features (second row) and when using both lexical and structural features (first row).

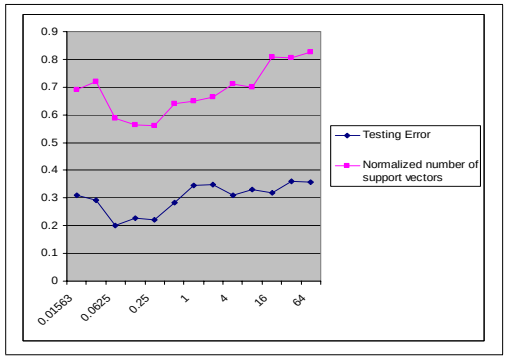


FIG. 2 – Plotting the error rates on testing data and the normalized number of support vectors when tuning γ, the RBF kernel parameter.

Figure 2 shows the influence of the kernel width both on the testing error curves and on the number of support vectors when $C^- = 10^{-2}$ (the optimal value selected through the procedure described in Section 4.1). We observe that in the optimality region the curve representing the error rates has a similar behavior as the curve corresponding to the normalized number of support vectors. That is, the minimum area in the number of support vectors corresponds to minimum error values of SVM-based topic segmentation on testing data. Therefore the number of support vectors is a good indicator of the optimality region.

We also observe from Figure 2 that the number of support vectors is rather large for all tuning values of γ. This reflects the fact that the positive samples (corresponding to the 'topic boundary' class) are not easily separable from the negative examples (corresponding to the 'non-topic boundary' class) due to noise. Moreover, our SVM approach has the critical property of differentiating between positive and negative class members by effectively removing the existing uninformative patterns from the data.

5 Comparison to other work

Comparing the performance of our model to other similar existing studies is not straightforward due to differences in corpora, in experimental design, and/or different input assumptions. Nevertheless, in the following we discuss some related work, by exemplifying some common aspects of the work and the experimental results.

Kauchak and Chen (2005) examined how the boundaries of thematic episodes can be detected in encyclopedia articles and in two books. They employ a supervised technique based on support vector machines using a variety of information including, for instance, features based on the presence of paragraph breaks, pronouns and named entities. When evaluating their topic segmentation model on encyclopedia articles, they obtained a P_k error rate of 39.8%.

Note that, in the context of spontaneous multiparty dialogue, the lack of paragraphs makes the topic segmentation task more difficult than the topic segmentation of narrative written text. For instance the chance of each paragraph break being a topic boundary is about 39.1% in expository texts (Hearst, 1997), while in the ICSI-MR corpus, the chance of each utterance to be a subtopic segment boundary is approximately 0.07% for top-level boundaries. Moreover, meeting dialogues provide particular challenges since topic changes are not always clearly delimited in contrast to e.g. broadcast news or written texts.

The model proposed in (Galley *et al.*, 2003) is the most similar to our model in terms of incorporating multi-party meeting specific features such as cue phrases, silences and conversation overlaps. Using such strucural features in addition to lexical chains, Galley et al. (2003) trained a decision tree which achieved a P_k error rate of 23% on a subset of the ICSI-MR corpus.

6 Conclusions and future work

In this article, we have presented an approach to learn the thematic structure of texts in the context of recorded and transcribed multi-party dialogs. Each utterance is represented as a collection of features obtained from lexical, syntactic and prosodic information. A SVM-based classifier has been trained to discriminate between utterances marking thematic and non-

thematic boundaries in meeting transcriptions.

Our contribution is fivefold. First, we introduce a series of different linguistic and acoustic cues to represent each utterance and we evaluate whether the proposed surface (meeting-specific) cues are useful for thematic segmentation. Second, we check the suitability of our SVM approach combining meeting-specific surface features with large-scale lexical features. Third, we evaluate the compatibility of SVM classification for various thresholds. Fourth, we study the influence of the kernel width on the testing error rate and on the (normalized) number of support vectors. Fifth, we compare the results with existing state-of-the-art methods for topic segmentation. We demonstrate that using 'surface' meeting specific features, our SVM approach generates competitive results on meeting data sets.

As a continuation of this work, it would be interesting to replicate our experiments on larger training sets. The proposed method can potentially be improved by exploiting additional sources of information, including for instance other prosodic information such as speech pitch range and speech rate. It would be also interesting to evaluate whether our topic segmentation approach can be further improved via other kernel methods.

Aknowledgments

This work is part of the Swiss National Center of Competence in Research on "Interactive Multimodal Information Management" (IM2, `http://www.im2.ch`), funded by the Swiss National Science Foundation.

References

BEEFERMAN D., BERGER A. & LAFFERTY J. (1999). Statistical Models for Text Segmentation. *Machine Learning*, **34**(Special Issue on Natural Language Learning), 177–210.

CHOI F. (2000). Advances in Domain Independent Linear Text Segmentation. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Seattle, USA.

GALLEY M., MCKEOWN K., FOSLER-LUISSIER E. & JING H. (2003). Discourse Segmentation of Multi-Party Conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 562–569, Sapporo, Japan.

GEORGESCUL M., CLARK A. & ARMSTRONG S. (2006a). An Analysis of Quantitative Aspects in the Evaluation of Thematic Segmentation Algorithms. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, p. 144–151, Sydney, Australia: Association for Computational Linguistics.

GEORGESCUL M., CLARK A. & ARMSTRONG S. (2006b). Word Distributions for Thematic Segmentation in a Support Vector Machine Approach. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*, p. 101–108, New York City, USA.

HEARST M. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, **23**(1), 33–64.

HIRSCHBERG J. & NAKATANI C. (1996). A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL)*, p. 286–293, Santa Cruz, California, USA.

JANIN A., ANG J., BHAGAT S., DHILLON R., EDWARDS J., MACIAS-GUARASA J., MORGAN N., PESKIN B., SHRIBERG E., STOLCKE A., WOOTERS C. & WREDE B. (2004). The ICSI Meeting Project: Resources and Research. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Meeting Recognition Workshop*, Montreal, Quebec, Canada.

JOACHIMS T. (1999). Making Large-Scale Support Vector Machine Learning Practical. In B. SCHÖLKOPF, C. BURGES & A. SMOLA, Eds., *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press.

KAUCHAK D. & CHEN F. (2005). Feature-Based Segmentation of Narrative Documents. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, p. 32–39, Ann Arbor, Michigan, USA.

LITMAN D. J. & PASSONNEAU R. J. (1995). Combining Multiple Knowledge Sources for Discourse Segmentation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 108–115, Cambridge, Massachusetts, USA.

MARCU D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press Cambridge, MA, USA.

PASSONNEAU R. J. & LITMAN D. J. (1993). Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proceedings of the 31st Conference on Association for Computational Linguistics (ACL)*, p. 148 – 155, Columbus, Ohio, USA.

PFAU T., ELLIS D. P. & STOLCKE A. (2001). Multispeaker Speech Activity Setection for the ICSI Meeting Recorder. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 107–110.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Stuttgart, Germany.

UTIYAMA M. & ISAHARA H. (2001). A Statistical Model for Domain-Independent Text Segmentation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL)*, p. 491–498, Toulouse, France.

VAPNIK V. N. (1998). *Statistical Learning Theory*. A Volume in the Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Berlin: Springer-Verlag.

Énergie textuelle de mémoires associatives

Silvia FERNÁNDEZ[1,2], Eric SANJUAN[1], Juan Manuel TORRES-MORENO[1,3]
[1] Laboratoire Informatique d'Avignon, BP 1228 84911 Avignon FRANCE
[2] LPM UHP-Nancy, BP 239 54506 Vandœuvre les Nancy FRANCE
[3] École Polytechnique de Montréal, CP 6079 Centre-ville, Montréal, Québec CANADA H3C3A7
{silvia.fernandez,eric.sanjuan,juan-manuel.torres}@univ-avignon.fr

Résumé. Dans cet article[1], nous présentons une approche de réseaux de neurones inspirée de la physique statistique de systèmes magnétiques pour étudier des problèmes fondamentaux du Traitement Automatique de la Langue Naturelle. L'algorithme modélise un document comme un système de neurones où l'on déduit l'énergie textuelle. Nous avons appliqué cette approche aux problèmes de résumé automatique et de détection de frontières thématiques. Les résultats sont très encourageants.

Abstract. In this paper we present a neural networks approach, inspired by statistical physics of magnetic systems, to study fundamental problems in Natural Language Processing. The algorithm models documents as neural network whose textual energy is studied. We obtained good results on the application of this method to automatic summarization and thematic borders detection.

Mots-clés : réseaux de neurones, réseaux de Hopfield, résumé, frontière thématiques.

Keywords: neural networks, Hopfield network, summarization, thematic boundary.

1 Introduction

Hopfield (Hopfield, 1982; Hertz *et al.*, 1991) s'est inspiré des systèmes physiques comme le modèle magnétique d'Ising (formalisme issu de la physique statistique decrivant un système avec des unités à deux états nommées spins) pour construire un réseau neuronal avec des capacités d'apprentissage et de récupération de patrons. Les capacités et limitations de ce réseau, appelé mémoire associative, ont été bien établies de façon théorique dans plusieurs études (Hopfield, 1982; Hertz *et al.*, 1991) : les patrons doivent être non corrélés afin que leur récupération soit sans erreur, le système sature rapidement et seulement une fraction des patrons peut être stockée correctement. Dès que leur nombre dépasse $\approx 0,14N$, aucun des patrons n'est plus reconnu. Cette situation restreint fortement leurs applications pratiques. Cependant, dans le cas du traitement automatique de la langue naturelle (TALN), nous pensons que l'on peut exploiter ce comportement. Le modèle vectoriel de textes (Salton & McGill, 1983), transforme les

[1]Ce travail a été realisé en partie grâce au financement du CONACYT (Mexico), bourse 175225.

phrases d'un document en vecteurs. Ces vecteurs peuvent être traités comme un réseaux de neurones type Hopfield. Si l'on définit un vocabulaire de taille N, où N est le nombre de termes uniques d'un document, on peut représenter une phrase comme une chaîne de N neurones actifs, $i = 1, \cdots, N$ (le mot i étant présent) ou inactifs (le mot i étant absent). Un document de P phrases, est composé de P chaînes dans l'espace vectoriel Ξ de dimension N. Ces vecteurs sont plus ou moins corrélés, selon les mots qu'ils partagent. Si les thématiques sont proches, il est raisonable de supposer que le degré de corrélation sera très élevé. Cela pose des problèmes si on essaie de stocker et de récupérer ces représentations dans un réseau type Hopfield. Cependant notre intérêt porte non pas sur la récupération, mais sur les interactions entre les mots et entre les phrases. Cette interaction nous allons la définir comme l'énergie textuelle d'un document. Elle peut servir, entre autres, à pondérer les phrases ou à détecter des changements entre des chaînes de neurones. Nous dévéloppons une métaphore qui permet d'utiliser le concept d'enérgie textuelle pour son application dans le résume générique ou la segmentation thématique. Nous présentons en Section 2 une brève introduction au modèle de Hopfield. En Section 3, nous faisons une extension de cette approche dans le traitement automatique de la langue naturelle. Nous utilisons ainsi des notions élémentaires de la théorie des graphes pour donner une interprétation de l'énergie textuelle comme une nouvelle mesure de similarité. En Section 4 nous appliquons nos algorithmes à la génération de résumés automatiques et à la détection de frontières thématiques, avant de conclure et presenter quelques perspectives.

2 L'approche énergétique de Hopfield

La contribution la plus importante de Hopfield à la théorie de réseaux de neurones a été l'introduction de la notion d'énergie issue de l'analogie avec les systèmes magnétiques. Un système magnétique est constitué d'un ensemble de N petits aimants appelés spins. Ces spins peuvent s'orienter selon plusieurs directions. Le cas le plus simple est représenté par le modèle d'Ising qui considére seulement deux directions possibles : vers le haut (↑, +1 ou 1) ou vers le bas (↓, -1 ou 0). Le modèle d'Ising a été utilisé dans une grande variété de systèmes qui peuvent être décrits par des variables binaires (Ma, 1985). Un système de N unités binaires possède $\nu = 1, ..., 2^N$ configurations (patrons) possibles. Dans le modèle de Hopfield les spins correspondent aux neurones qui interagissent selon la règle d'apprentissage d'Hebb[2] :

$$J^{i,j} = \sum_{\mu=1}^{P} s^i_\mu s^j_\mu \tag{1}$$

s^i et s^j sont les états des neurones i et j. Les autocorrelations ne sont pas calculées ($i \neq j$). La sommation porte sur les P patrons a stocker. Cette règle d'interaction est locale, car $J^{i,j}$ dépend seulement des états des unités connectées. Ce modèle est connu aussi comme mémoire associative. Elle possède la capacité de stocker et de récupérer un certain nombre de configurations du système, car la règle de Hebb transforme ces configurations en attracteurs (minimaux locaux) de la fonction d'énergie (Hopfield, 1982) :

$$E = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} s^i \; J^{i,j} \; s^j \tag{2}$$

[2]Hebb (Hertz *et al.*, 1991) a suggéré que les connexions synaptiques changent proportionnellement à la corrélation entre les états des neurones.

L'énergie est fonction de la configuration du système, c'est-à-dire, de l'état (d'activation ou non activation) de toutes ces unités. Si on présente un patron ν, chaque spin subira un champ local $h^i = \sum_{j=1}^{N} J^{i,j} s^j$ induit par les autres N spins (voir figure 1). Les spins s'aligneront selon h^i

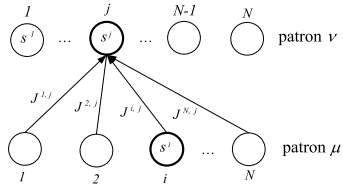


FIG. 1 – Champ h_i subi par le spin $s_j, \in$ la chaîne (patron) ν produit par les autres N spins $\in \mu$.

pour restituer le patron stocké qui est le plus proche au patron présenté ν. Nous n'allons pas détailler la méthode de recupération de patrons[3], car notre intérêt va porter sur la distribution et les propriétés de l'énergie du système (2). Cette fonction monotone et décroissante avait été utilisée uniquement pour montrer que l'apprentissage est borné. D'un autre côté, le modèle vectoriel (Salton & McGill, 1983) transforme un document dans un espace adéquat où une matrice S contient l'information du texte sous forme de sacs de mots. On peut considérer S comme l'ensemble des configurations d'un système dont on peut calculer l'énergie.

3 Applications au TALN

Les documents sont pré-traités avec des algorithmes classiques de filtrage de mots fonctionnels[4], de normalisation et de lemmatisation (Porter, 1980; Manning & Schutze, 2000) afin de réduire la dimensionnalité. Une représentation en sac de mots produit une matrice $S_{[P\times N]}$ de fréquences/absences composée de $\mu = 1, \cdots, P$ phrases (lignes) ; $\vec{\sigma}_\mu = \{s^1_\mu, \cdots, s^i_\mu, \cdots, s^N_\mu\}$ et un vocabulaire de $i = 1, \cdots, N$ termes (colonnes).

$$S = \begin{pmatrix} s^1_1 & s^2_1 & \cdots & s^N_1 \\ s^1_2 & s^2_2 & \cdots & s^N_2 \\ \vdots & \vdots & \ddots & \vdots \\ s^1_P & s^2_P & \cdots & s^N_P \end{pmatrix} \; ; \; s^i_\mu = \begin{cases} TF^i & \text{si le terme } i \text{ existe} \\ 0 & \text{autrement} \end{cases} \tag{3}$$

La présence du mot i représente un spin $s^i \uparrow$ avec une magnitude donnée par sa fréquence TF^i (son absence par $\downarrow$ respectivement), et une phrase $\vec{\sigma}_\mu$ est donc une chaîne de N spins. Nous allons nous différencier de (Hopfield, 1982) sur deux points : S est une matrice entière (ses éléments prennent des valeurs fréquentielles absolues) et nous utilisons les éléments $J^{i,i}$ car cette auto-corrélation permet d'établir l'interaction du mot i parmi les P phrases, ce qui est important en TALN. Pour calculer les interations entre les N termes du vocabulaire, on applique la règle de correlation de Hebb, qui en forme matricielle est égale à :

$$J = S^T \times S \tag{4}$$

[3]Cependant le lecteur intéressé peut consulter, par exemple (Hopfield, 1982; Kosko, 1988; Hertz *et al.*, 1991).
[4]Nous avons effectué le filtrage de chiffres et l'utilisation d'anti-dictionnaires.

Chaque élément $J^{i,j} \in J_{[N\times N]}$ est équivalent au calcul de (1). L'énergie textuelle d'interaction (2) peut alors s'exprimer comme :

$$E = -\frac{1}{2} S \times J \times S^T \tag{5}$$

Un élément $E_{\mu,\nu} \in E_{[P\times P]}$ représente l'énergie d'interaction entre les patrons μ et ν (figure 1).

3.1 L'énergie textuelle : une nouvelle mesure de similarité

Nous allons expliquer théoriquement la nature des liens entre phrases que la mesure d'énergie textuelle induit. Pour cela nous utilisons quelques notions élémentaires de la théorie des graphes. L'interprétation que nous allons faire repose sur le fait que la matrice (5) peut s'écrire :

$$E = -\frac{1}{2} S \times (S^T \times S) \times S^T = -\frac{1}{2}(S \times S^T)^2 \tag{6}$$

Considérons les phrases comme des ensembles σ de mots. Ces ensembles constituent les sommets du graphe. On trace une arête entre deux de ces sommets σ_μ, σ_ν chaque fois qu'ils partagent au moins un mot en commun $\sigma_\mu \cap \sigma_\nu \neq \emptyset$. On obtient ainsi le graphe $I(S)$ d'intersection des phrases (voir un exemple à quatre phrases en figure 2). On value ces paires $\{\sigma_1, \sigma_2\}$ que l'on appelle arêtes par le nombre exact $|\sigma_1 \cap \sigma_2|$ de mots que partagent les deux sommets reliés. Enfin, on ajoute à chaque sommet σ une arête de réflexivité $\{\sigma\}$ valuée par le cardinal $|\sigma|$ de σ. Ce graphe d'intersection valué est isomorphe au graphe $G(S \times S^T)$ d'adjacence de la matrice carrée $S \times S^T$. En effet, $G(S \times S^T)$ contient P sommets. Il existe une arête entre deux sommets μ, ν si et seulement si $[S \times S^T]_{\mu,\nu} > 0$. Si c'est le cas, cette arête est valuée par $[S \times S^T]_{\mu,\nu}$, valeur qui correspond au nombre de mots en commun entre les phrases μ et ν. Chaque sommet μ est pondéré par $[S \times S^T]_{\mu,\mu}$ ce qui correspond à l'ajout d'une arête de reflexivité. Il en résulte

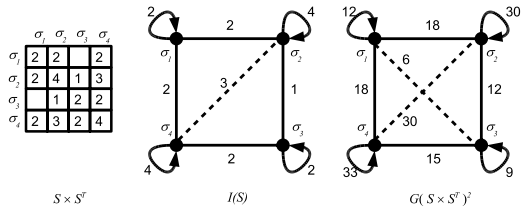


FIG. 2 – Graphes d'adjacence issus de la matrice d'énergie.

que la matrice d'énergie textuelle E est la matrice d'adjacence du graphe $G(S \times S^T)^2$ dont :
- les sommets sont les mêmes que ceux du graphe d'intersection $I(S)$;
- il existe une arête entre deux sommets chaque fois qu'il existe un chemin de longueur au plus 2 dans le graphe d'intersection ;
- la valeur d'une arête : a) boucle sur un sommet σ est la somme des carrés des valeurs des arêtes adjacentes au sommet et b) entre deux sommets distincts σ_μ et σ_ν adjacents est la somme des produits des valeurs des arêtes sur tout chemin de longueur 2 entre les deux sommets. Ces chemins pouvant comprendre des boucles.

De cette représentation on en déduit que la matrice d'énergie textuelle relie à la fois des phrases ayant des mots communs puisque elle englobe le graphe d'intersection, ainsi que des phrases qui partagent un même voisinage sans pour autant partager nécessairement un même vocabulaire. C'est à dire que deux phrases σ_1, σ_3 ne partageant aucun mot en commun mais pour lesquelles il existe au moins une troisième phrase σ_2 telle que $\sigma_1 \cap \sigma_2 \neq \emptyset$ et $\sigma_3 \cap \sigma_2 \neq \emptyset$ seront tout de même reliées. La force de ce lien dépend premièrement du nombre de phrases σ_2 dans leur voisinage commun, et donc du vocabulaire apparaissant dans un contexte commun.

4 Expériences et résultats

L'énergie textuelle peut être utilisée comme mesure de similarité dans les applications du TALN. De façon intuitive, cette similarité peut servir à scorer les phrases d'un document et séparer ainsi celles qui sont pertinentes de celles qui ne le sont pas. Ceci conduit immediatement à une stratégie de résumé automatique par extraction de phrases. Une autre approche, moins évidente, consiste à utiliser l'information de cette énergie (vue comme un spectre ou signal numérique de la phrase) et de la comparer au spectre de toutes les autres. Un test statistique peut alors indiquer si ce signal est semblable à celui d'autres phrases regroupés en segments ou pas. Ceci peut être vu comme une détection de frontières thématiques dans un document.

4.1 Résumé automatique

Sous l'hypothèse que l'énergie d'une phrase μ reflète son poids dans le document, nous avons appliqué (6) au résumé par extraction de phrases (Mani & Maybury, 1999; Radev *et al.*, 2002). Cette méthode est orientée, pour le moment, à la génération de résumés génériques monodocument. Cependant, nous pensons qu'une modification de l'approche (voir Section 5) pourrait nous permettre d'obtenir des résumés guidés par une requête ou un sujet défini par l'utilisateur (ce qui correspond au protocole des conféréences DUC[5]). L'algorithme de résumé comprend trois modules. Le premier réalise la transformation vectorielle du texte avec des processus de filtrage, de lemmatisation/*stemming* et de normalisation. Le second module applique le modèle de spins et réalise le calcul de la matrice d'énergie textuelle (6). Nous obtenons la pondération de la phrase ν en utilisant ses valeurs absolues d'énergie, c'est-à-dire, en triant selon $\sum_\mu |\boldsymbol{E}_{\mu,\nu}|$. Ainsi, les phrases pertinentes seront selectionnées comme ayant la plus grande énergie absolue. Finalement, le troisième module génère les résumés par affichage et concaténation des phrases pertinentes. Les deux premiers modules reposent sur le système Cortex[6]. Pour les tests en français[7] nous avons choisi les textes : « 3-mélanges » composé de trois thématiques, « puces » de deux thématiques et « J'accuse » (lettre d'Émile Zola). Deux textes de la wikipedia en anglais ont été analysés, « Lewinksky » et « Québec »[8]. Nous avons evalué les résumés produits par notre système avec ROUGE (Lin, 2004), qui mesure la similarité, suivant plusieurs stratégies, entre un résumé candidat (produit automatiquement) et des résumés de référence (créés par des humains). Nous comparons dans les tables 1 à 5 les performances de la méthode d'énergie, de

[5]Document Understandig Conferences http ://www-nlpir.nist.gov/projects/duc/index.html

[6]Le système Cortex (Torres-Moreno *et al.*, 2002) effectue une extraction non supervisée de phrases pertinentes en utilisant plusieurs métriques pilotées par un algorithme de décision.

[7]Recupérables à l'adresse http ://www.lia.univ-avignon.fr.

[8]http ://en.wikipedia.org/wiki/Monica_Lewinsky, http ://en.wikipedia.org/wiki/Quebec_sovereignty_movement

Cortex et d'une *baseline* où les phrases ont été choisies au hasard. Nous constatons que notre méthode est comparable au système Cortex en termes de précision, de rappel et de F-score.

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,49577	0,50635	**0,49676**	**0,50643**	0,29125	0,3117
Précision	**0,43229**	**0,44114**	0.42288	0,43068	0,32801	0,35191
F**-score**	**0,46186**	**0,47150**	0.45685	0,46549	0,30744	0,32936

TAB. 1 – Texte « 3-mélanges » (27 phrases, 826 mots ; résumé au 25% ; 8 résumés référence).

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,52040	0,53353	**0,53595**	**0,55878**	0,25938	0,27721
Précision	0,52469	0,53796	**0,53120**	**0,55380**	0,37589	0,40474
F**-score**	0,52254	0,53574	**0,53356**	**0,55628**	0,30530	0,32723

TAB. 2 – Texte « puces » (29 phrases, 653 mots ; résumé au 25% ; 8 résumés référence).

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,61457	0,64192	**0,63160**	**0,65987**	0,18690	0,20185
Précision	0,51425	0,53700	**0,52725**	**0,55071**	0,30920	0,37195
F**-score**	0,55995	0,58479	**0,57473**	**0,60037**	0,21766	0,26152

TAB. 3 – Texte « J'accuse » (206 phrases, 4936 mots ; résumé au 12% ; 6 résumés référence).

4.2 Détection de frontières thématiques

Plusieurs stratégies ont été développées pour segmenter thématiquement un texte. Elles peuvent être supervisées ou non. On trouve PLSA (Brants *et al.*, 2002) qui estime les probabilités d'appartenance des termes à des classes sémantiques, des méthodes s'appuyant sur des modèles de Markov (Amini *et al.*, 2000), sur une classification des termes (Caillet *et al.*, 2004; Chuang & Chien, 2004) ou sur des chaînes lexicales (Sitbon & Bellot, 2005). De façon originale, nous avons utilisé la matrice d'énergie E (6). Ce choix permet de s'adapter à de nouvelles thématiques et de rester indépendant vis à vis de la langue des documents. Pour pouvoir comparer les énergies entre elles nous introduisons le coefficient de concordance W de Kendall (Siegel & Castellan, 1988) et le calcul de sa $p-$valeur. Ils permettent de définir un test statistique de concordance entre k juges qui classent un ensemble de P objets. Nous avons utilisé ce test pour trouver les frontières thématiques entre segments. Nous montrons en figure (3) l'énergie d'interaction entre quelques phrases d'un texte composé de deux thématiques. Étant donné que (6) est capable de détecter et de pondérer le voisinage d'une phrase, on peut constater une similarité entre les courbes de l'une (gras) et de l'autre thématique (pointillé). Voici le protocole de test que nous avons adopté.

1. Selon la nature du texte (homogène ou hétéroclite) on émet a priori l'une des deux hypothèses initiales H_0 qui suivent : *i)* la phrase $\mu + 1$ appartient à la même thématique que la phrase précédente μ ou au contraire *ii)* la phrase $\mu + 1$ marque une rupture avec μ.

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,56107	0,57859	**0,61832**	**0,62705**	0,24227	0,25584
Précision	0,39516	0,40658	**0,42587**	**0,43085**	0,32490	0,34393
F**-score**	0,46372	0,47757	**0,50436**	**0,51076**	0,27671	0,29248

TAB. 4 – Texte « Lewinsky » (30 phrases, 816 mots ; résumé au 20% ; 7 résumés référence).

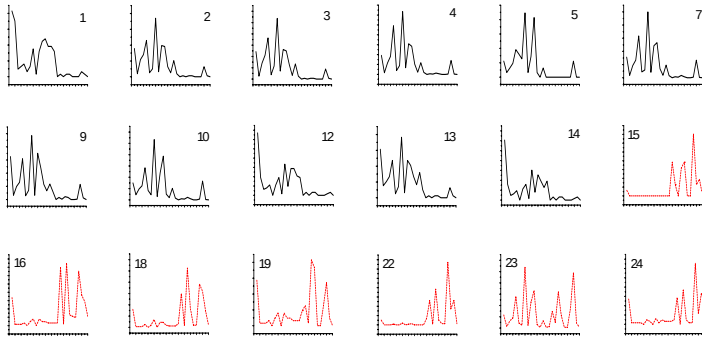


FIG. 3 – Énergie textuelle de « 2-mélanges ». En trait continu l'énergie des phrases de la 1ère thématique, en pointillé celle de la 2ème. Le changement d'allure des courbes entre les phrases 14-15 correspond à un changement thématique. L'axe horizontal indique le numéro de phrase dans l'ordre du document. L'axe vertical, l'énergie textuelle de la phrase affichée vs. les autres.

2. On estime alors la probabilité p que l'hypothèse H_0 choisie soit vérifiée en calculant le coefficient de concordance W de Kendall entre les deux classements par proximité induits par les phrases μ et $\mu + 1$ sur les autres phrases. Le coefficient W de Kendall vaut 1 en cas d'accord total entre les classements et 0 dans la cas de désaccord total. La probabilité p est alors estimée en utilisant l'approximation de la loi du W par une loi du χ^2.
3. Finalement, si $p < 0,1$ on rejette H_0 et l'on adopte l'hypothèse alternative (son complémentaire) avec un risque p de se tromper. Il est important de préciser que la valeur seuil $0,1$ est fixée a priori conformémént à la méthodologie statistique inférentielle.

Il s'agit donc de tests non-paramétriques qui ne requièrent aucune supposition sur une éventuelle distribution gaussienne des données. Pour chaque document, nous avons éliminé les phrases dont l'énergie est inférieure à un seuil. Ces phrases sont celles qui contiennent un nombre de mots < 2 (patrons à spins 0) ou trop longues (si l'on a suffisamment de phrases par segment), et qui induisent un fort bruit dans E. Les figures (4) et (5) montrent la détection

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,50945	0,53773	**0,56364**	**0,58716**	0,27344	0,32127
Précision	0,46276	0,48824	**0,50803**	**0,52899**	0,33254	0,39092
F**-score**	0,48498	0,51179	**0,53439**	**0,55656**	0,29991	0,35244

TAB. 5 – Texte « Québec » (44 phrases, 1190 mots ; résumé au 25% ; 8 résumés référence).

des frontières pour les textes à 2 et 3 thématiques. Les véritables frontières sont indiquées en pointillé. Ce protocole de test, en adoptant l'hypothèse *ii)* comme H_0, a détecté une frontière entre les phrases 14-15 pour le texte « 2-mélanges ». Pour le texte « 3-mélanges », le test a trouvé deux frontières entre les segments 8-9 et 16-18. Dans les deux cas, cela correspond effectivement aux frontières thématiques. Une troisième (fausse) frontière a été signalée entre les phrases 23-24 du texte « 2-mélanges ». Cela mérite d'être commenté : si on regarde sur la figure (3) l'énergie de la phrase 23, elle est bien différente de celle des phrases 22 ou 24. La phrase 23 présente une courbe chevauchant les deux thématiques. C'est pourquoi le test ne peut pas l'identifier comme appartenant à la même classe. Le même raisonnemment tient pour toutes les fausses frontières. Pour le texte « physique-climat-chanel » le test du W de Kendall a détecté

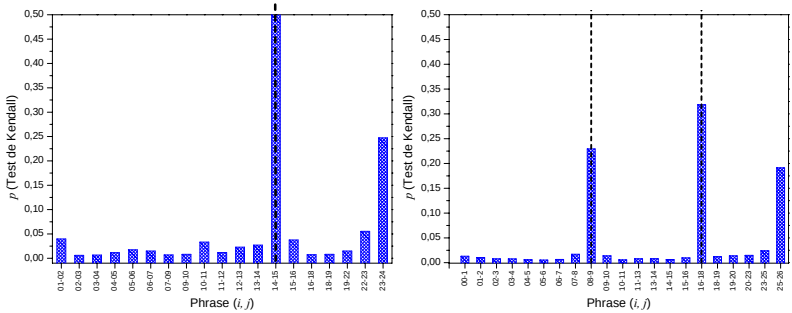


FIG. 4 – Détection des frontières pour le texte « 2-mélanges » (2 thématiques, à gauche) et « 3-mélanges » (3 thématiques, à droite).

trois frontières entre les phrases 5-6 et 12-15, qui correspondent aux frontières effectives. Pour le texte en anglais à deux thématiques le test a trouvé une frontière entre les segments 44-45 qui correspond à la vraie frontière. Nous avons calculé le F-score de façon similaire à DEFT 2005

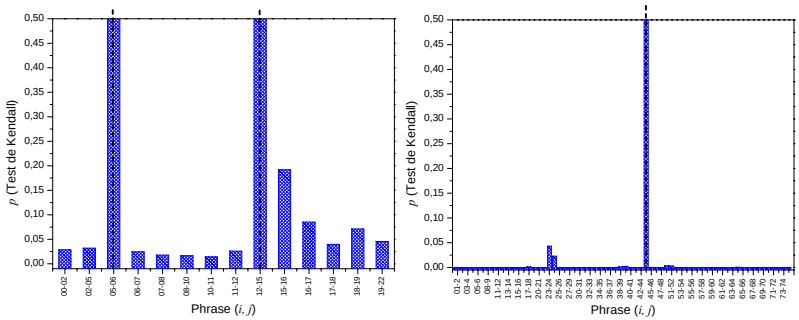


FIG. 5 – Détection des frontières pour le texte en français à 3 thématiques « physique-climat-chanel » à gauche et en anglais « québec-lewinsky » à droite.

(Alphonse *et al.*, 2005) [9]. Ainsi pour « 2-mélanges » F-score = 0,66 ; « 3-mélanges » F-score =

[9] En considérant $\beta = 1$, $F\text{-score}(\beta) = \frac{2 \times \text{Nb\_frontières\_correctes\_extraites}}{\text{Nb\_total\_frontières\_extraites} + \text{Nb\_total\_véritables}}$

0,66 ; « physique-climat-chanel » F-score = 0,80 et « québec-lewinsky » F-score = 1. Dans une autre expérience, nous avons comparé notre système à deux autres : LCseg (Galley *et al.*,) et LIA_seg (Sitbon & Bellot, 2005), qui utilisent tous les deux des chaînes lexicales. Le corpus de reférence a été construit à partir d'articles du journal Le Monde. Il est composé d'ensembles de 100 documents où chacun correspond à la taille moyenne des segments pré-définie. Un document est composé de 10 segments extraits d'articles thématiquement différentes tirés au hasard. Compte tenu des particularités de ce corpus nous avons adopté *i* comme hypothèse initiale H_0. Les scores sont calculés avec Windiff (Pevzner & Hearst, 2002), utilisée dans la segmentation thématique. Cette fonction mesure la différence entre les frontières véritables et celles trouvées automatiquement dans une fenêtre glissante : plus la valeur est petite, plus le système est performant. LIA_seg depend d'un paramètre qui donne lieu à différentes performances (d'où la plage de valeurs affichée). Notre méthode obtient des performances comparables aux systèmes dans l'état de l'art mais en utilisant bien moins de paramètres, en particulier nous ne faisons aucune supposition sur le nombre de thématiques à détecter.

Taille du segment (en phrases)	**LCseg**	**LIA_seg**	**Énergie**
9-11	0,3272	(**0,3187**-0,4635)	0,4419
3-11	0,3837	(**0,3685**-0,5105)	0,4403
3-5	0,4344	(0,4204-0,5856)	**0,4167**

TAB. 6 – Mesure Windiff pour LCseg, LIA_seg et Énergie (segments de différentes tailles).

5 Conclusion et perspectives

Nous avons introduit le concept d'énergie textuelle basé sur des approches des réseaux de neurones. Cela nous a permis de développer un nouvel algorithme de résumé automatique. Des tests effectués ont montré que notre algorithme est adapté à la recherche de segments pertinents. On obtient des résumés équilibrés où la plupart des thèmes sont abordés dans le condensé final. Les avantages supplémentaires consistent à ce que les résumés sont obtenus de façon indépendante de la taille du texte, des sujets abordés, d'une certaine quantité de bruit et de la langue (sauf pour la partie pré-traitement). Nous pensons que l'algorithme d'énergie pourrait être incorporé au système Cortex, où il jouerait le rôle d'une des métriques pilotée par un algorithme de décision. Ceci permettrait d'obtenir des résumés à l'aide d'une requête de l'utilisateur ou des résumés multi-documents. Une autre voie intéressante est le calcul des propriétés comme la « magnétisation » d'un document. (Shukla, 1997) a etudié des phénomènes magnétiques dans les réseaux de neurones type Hopfield dont on pense se servir. On étudiera la réponse du système face à l'application d'un champ externe. Ce champ, représenté par le vecteur des termes d'un texte decrivant une thématique (topique) sera mis en relation avec un document. Ainsi, les phrases du document pourraient, ou non, s'aligner selon leur degré de pertinence par rapport à la thématique. Ceci permettrait de générer des résumés personnalisés, telles que définis dans les tâches DUC. Nous avons également abordé le problème de la détection des frontières thématiques des documents. La méthode, basée sur la matrice d'énergie du système de spins, est couplée à un test statistique non-paramètrique robuste. Les résultats sont très encourageants. Une critique de la méthode d'énergie pourrait être qu'elle nécessite la manipulation (produits, transposée) d'une matrice de taille $[P \times P]$. Cependant la représentation sous forme de graphe nous permet d'exécuter ces calcus en temps $P \log(P)$ et en espace P^2.

Références

ALPHONSE E., AMRANI A., AZÉ J., HEITZ T., MEZAOUR A.-D. & ROCHE M. (2005). Préparation des données et analyse des résultats de DEFT'05. In *TALN 2005 - Atelier DEFT'05*, volume 2, p. 95–97.

AMINI M.-R., ZARAGOZA H. & GALLINARI P. (2000). Learning for sequence extraction tasks. In *RIAO 2000*, p. 476–489.

BRANTS T., CHEN F. & TSOCHANTARIDIS I. (2002). Topic-based document segmentation with probabilistic latent semantic anaysis. In *CIKM'02*, p. 211–218, McLean, Virginia, USA.

CAILLET M., PESSIOT J.-F., AMINI M. & GALLINARI P. (2004). Unsupervised learning with term clustering for thematic segmentation of texts. In *RIAO'04*, p. 648–657, France.

CHUANG S.-L. & CHIEN L.-F. (2004). A practical web-based approach to generating Topic hierarchy for Text segments. In *Thirteenth ACM conference on Information and knowledge management*, p. 127–136, Washington, D.C., USA.

GALLEY M., MCKEOWN K. R., FOSLER-LUSSIER E. & JING H. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–03)*, Sapporo, Japan.

HERTZ J., KROGH A. & PALMER G. (1991). *Introduction to the theorie of Neural Computation*. Redwood City, CA : Addison Wesley.

HOPFIELD J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, **9**, 2554–2558.

KOSKO B. (1988). Bidirectional associative memories. *IEEE Transactions Systems Man, Cybernetics*, **18**, 49–60.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out (WAS 2004)*.

MA S. (1985). *Statistical Mechanics*. Philadelphia, CA : World Scientific.

MANI I. & MAYBURY M. T. (1999). *Advances in Automatic Text Summarization*. MIT Press.

MANNING C. D. & SCHUTZE H. (2000). *Foundations of Statistical Natural Language Processing*. The MIT Press.

PEVZNER L. & HEARST M. (2002). A critique and improvement of an evaluation metric for text segmentation. In *Computational Linguistic*, volume 1, p. 19–36.

PORTER M. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

RADEV D., WINKEL A. & TOPPER M. (2002). Multi Document Centroid-based Text Summarization. In *ACL 2002*.

SALTON G. & MCGILL M. (1983). *Introduction to modern information retrieval*. Computer Science Series McGraw Hill Publishing Company.

SHUKLA P. (1997). Response of the Hopfield-Little model in an applied field. *Physical Review E*, **56**(2), 2265–2268.

SIEGEL S. & CASTELLAN N. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw Hill.

SITBON L. & BELLOT P. (2005). Segmentation thématique par chaînes lexicales pondérées. In *TALN 2005*, volume 1, p. 505–510.

TORRES-MORENO J.-M., VELÁZQUEZ-MORALES P. & MEUNIER J. (2002). Condensés de textes par des méthodes numériques. In *JADT*, volume 2, p. 723–734.

Session
Acquisition

Une expérience d'extraction de relations sémantiques à partir de textes dans le domaine médical

Mehdi EMBAREK, Olivier FERRET
CEA LIST, LIC2M,
18 route du Panorama, BP6, FONTENAY AUX ROSES, F- 92265 France
{embarekm,ferreto}@zoe.cea.fr

Résumé. Dans cet article, nous présentons une méthode permettant d'extraire à partir de textes des relations sémantiques dans le domaine médical en utilisant des patrons linguistiques. La première partie de cette méthode consiste à identifier les entités entre lesquelles les relations visées interviennent, en l'occurrence les maladies, les examens, les médicaments et les symptômes. La présence d'une des relations sémantiques visées dans les phrases contenant un couple de ces entités est ensuite validée par l'application de patrons linguistiques préalablement appris de manière automatique à partir d'un corpus annoté. Nous rendons compte de l'évaluation de cette méthode sur un corpus en Français pour quatre relations.

Abstract. In this article, we present a method to extract semantic relations automatically in the medical domain using linguistic patterns. This method consists first in identifying the entities that are part of the relations to extract, that is to say diseases, exams, treatments, drugs and symptoms. Thereafter, sentences that contain these entities are extracted and the presence of a semantic relation is validated by applying linguistic patterns that were automatically learnt from an annotated corpus. We report the results of an evaluation of our extraction method on a French corpus for four relations.

Mots-clés : extraction de relations sémantiques, patrons lexico-syntaxiques, domaine médical.

Keywords: extraction of semantic relations, lexico-syntactic patterns, medical domain.

1 Introduction

Dans cet article, nous nous intéressons au domaine médical, dont la particularité est la richesse et la complexité de son vocabulaire spécialisé. Cette particularité a conduit depuis de nombreuses années au développement d'un ensemble important de ressources terminologiques telles que le MeSH ou l'UMLS par exemple. Ces ressources ont été utilisées dans des contextes aussi divers que l'indexation de documents, la recherche d'information, l'extraction d'information ou même les systèmes de question-réponse. À l'image de réseaux lexicaux de même type mais plus généraux, comme WordNet (Fellbaum, 1998), ces ressources contiennent majoritairement des relations d'hyperonymie ou de synonymie et sont donc beaucoup moins riches en relations syntagmatiques comme celles spécifiant qu'une maladie *M* peut être soignée par le traitement *T* ou que l'examen *E* permet de diagnostiquer la maladie *M*. De même, les méthodes ayant pour

objectif d'extraire des relations sémantiques à partir de textes se focalisent majoritairement sur les relations de synonymie et d'hyperonymie, à la suite de (Hearst, 1992) ou plus récemment de (Caraballo, 1999).

Le travail dont nous rendons compte dans cet article concerne l'extraction et la validation de relations sémantiques entre des entités caractéristiques du domaine médical, telles que des maladies, des médicaments ou des examens, en se focalisant prioritairement sur des relations de type syntagmatique. Différents travaux ont déjà été menés concernant l'extraction de relations sémantiques dans le domaine médical ou biomédical, travaux parmi lesquels on peut citer (Craven, 1999), (Mukherjea & Sahay, 2006), (Rosario & Hearst, 2004) ou (Vintar & Buitelaar, 2003). Les recherches menées en extraction d'information dans ce même contexte, bien qu'ayant *a priori* une finalité plus large, se ramènent dans bon nombre de cas à l'extraction de ce même type de relations, à l'instar de la détection des interactions entre gènes ou entre gènes et protéines. On se reportera à (Nédellec, 2004) pour un panorama de ces travaux, souvent fondés sur des règles d'extraction définies manuellement.

La méthode que nous proposons repose pour sa part sur l'identification puis l'application de patrons linguistiques caractérisant les relations visées, dans le prolongement direct de (Pantel *et al.*, 2004). Cette application se déroule en deux étapes. La première consiste à identifier dans les textes les entités du domaine médical intervenant dans les relations visées. Dans la phrase « … en novembre 2001, année d'un cancer de la prostate traité par radiothérapie et qu'il affirme aujourd'hui disparu, … », le premier objectif est ainsi de repérer que *cancer de la prostate* est une maladie et que *radiothérapie* est un traitement. Dans un second temps, l'application du patron <*maladie*> *traité par* <*traitement*> construit automatiquement à partir d'un corpus de référence permet de valider la présence d'une relation entre ces deux entités, relation stipulant dans le cas présent que la *radiothérapie* est un traitement possible du *cancer de la prostate*.

2 Ontologie du domaine médical

La première étape de notre travail s'est focalisée sur la définition d'une ontologie du domaine de la médecine générale permettant de faire apparaître les entités caractérisant ce domaine ainsi que les relations existant entre ces entités. Cette ontologie a été définie à la fois en sollicitant directement des médecins et par l'analyse des questions typiquement posées par des médecins généralistes (Ely *et al.*, 1999). La Figure 1 illustre le sous-ensemble de cette ontologie correspondant aux quatre relations objets de notre travail et aux entités qu'elles font intervenir. Sont

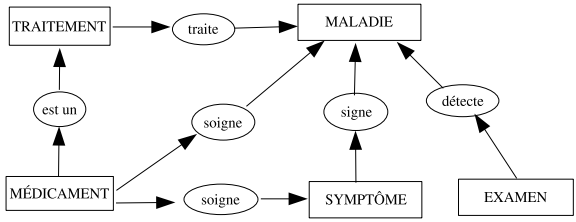


FIG. 1 – Sous-ensemble de l'ontologie du domaine médical concernant les relations à extraire

ainsi concernées les cinq entités suivantes : Maladie, Traitement, Médicament, Symptôme et Examen. Parmi toutes les relations de la Figure 1, les quatre relations auxquelles nous nous sommes attachés sont :

– Traite : Maladie – Traitement
– Soigne : Maladie – Médicament
– Détecte : Maladie – Examen
– Signe : Maladie – Symptôme

3 Reconnaissance des entités médicales

L'identification dans les documents des entités médicales que nous avons retenues a été réalisée indépendamment de leur nature intrinsèque (entité nommée au sens littéral ou terme) en adoptant une approche à base de règles mêlant patrons morpho-syntaxiques et listes d'entités ou d'éléments caractéristiques de ces entités. Ces règles ont été définies manuellement à partir d'un travail sur corpus. Nous avons repris en l'occurrence une des approches classiquement utilisées pour identifier des entités nommées de nature plus générale comme les personnes, les organisations ou les lieux. À la différence de ces dernières, les patrons morpho-syntaxiques ont ici une importance moindre ce qui à l'inverse, donne un rôle plus central aux listes d'entités ou de parties d'entités. Nous avons donc accordé un soin tout particulier à la constitution de ces listes pour chaque type d'entités en exploitant pour ce faire à la fois des ressources disponibles sur le Web[1] et des dictionnaires de l'Académie de Médecine sous forme électronique.

Chaque règle de reconnaissance d'une entité est composée d'un déclencheur, d'un contexte précédent, d'un contexte suivant et du type d'entité identifié. Ces règles sont implémentées sous la forme d'automates. Leur application s'effectuant à la suite de l'étiquetage morpho-syntaxique réalisé par l'analyseur LIMA (LIC2M Multilingual Analyzer) (Besançon & Chalendar, 2005), le déclencheur et les contextes précédent et suivant d'une règle prennent donc la forme d'expressions régulières pouvant porter sur la forme fléchie, la forme normalisée ou la catégorie d'un ou de plusieurs mots. Ainsi, la règle

@AnnonceurMaladie::$L_DET ?::$L_DET ($L_NC|$L_NP)::MALADIE[2]
déclencheur::*contexte_précédent*::*contexte_suivant*::*type_d'expression*

permet-elle d'identifier *maladie de Lyme* comme une maladie dans « La maladie de Lyme est une … » tandis que la règle

[@AnnonceurSymptome]::::[,] [$L_NC] [$L_DET] $L_NC::SYMPTOME[3]

reconnaît *fièvre* comme un symptôme dans « … symptôme, comme la fièvre … ». On peut noter à cette occasion la présence de référence à des listes permettant de regrouper des éléments linguistiques ayant un même rôle, comme les éléments marquant la présence d'une maladie (*@AnnonceurMaladie={maladie, syndrome …}*) ou ceux marquant la présence d'un symptôme (*@AnnonceurSymptome={signe, symptôme …}*).

[1] Le site Doctissimo pour les noms de médicaments ou le site Orphanet pour les noms de maladies par exemple.
[2] ? marque classiquement un élément optionnel tandis que (|) note une alternative. *$L_DET*, *$L_NC* et *$L_NP* sont des catégories morpho-syntaxiques, correspondant respectivement à déterminant, nom commun et nom propre.
[3] [] permet de spécifier la non appartenance d'un élément à l'entité reconnue.

4 Extraction de relations sémantiques

4.1 Apprentissage de patrons linguistiques d'extraction

Le terme de patron linguistique désigne dans le cas présent un schéma lexico-syntaxique spécifique d'une relation sémantique intervenant entre deux entités. Dans le cas présent, ces patrons sont dits multi-niveaux, c'est-à-dire qu'ils s'appuient sur des informations provenant de plusieurs niveaux de traitement des textes : à l'instar des règles de reconnaissance des entités médicales, ils peuvent ainsi faire intervenir la forme fléchie des mots, leur forme normalisée ou bien encore leur catégorie morpho-syntaxique. Le processus que nous avons élaboré pour extraire à partir d'un corpus les patrons linguistiques caractérisant une relation est le suivant :

1. appliquer sur le corpus considéré les règles de reconnaissance des entités médicales impliquées dans la relation cible. Nous prendrons à titre d'exemple la relation *Traite* entre une *Maladie* et un *Traitement* ;
2. extraire du corpus toutes les phrases contenant les deux entités de la relation cible, à savoir ici les phrases contenant à la fois une maladie et un traitement ;
3. sélectionner manuellement les phrases dans lesquelles la relation entre les deux entités correspond effectivement à la relation cible. Cela implique en particulier d'écarter les phrases telles que « la <maladie> n'est pas traitée par le <traitement> » ;
4. réaliser l'analyse linguistique de chaque phrase sélectionnée pour en faire apparaître les différents niveaux d'information. Cette analyse est réalisée comme pour la reconnaissance des entités par l'analyseur LIMA ;
5. remplacer dans chaque phrase les entités par leur type ;
6. appliquer l'algorithme d'extraction de patrons multi-niveaux présenté ci-dessous entre chaque couple de phrases parmi celles sélectionnées précédemment.

Pour extraire les patrons linguistiques propres à chaque relation sémantique traitée (cf. Section 2), nous faisons appel à l'algorithme proposé par Ravichandran dans (Pantel *et al.*, 2004) pour apprendre des patrons multi-niveaux. Cet algorithme est composé de deux parties. La première consiste à calculer la distance d'édition minimale entre deux phrases, ce qui permet de déterminer le nombre minimum d'opérations (insertion, suppression et remplacement) à appliquer pour passer d'une phrase à l'autre. La deuxième étape extrait le patron multi-niveau le plus spécifique permettant de généraliser les deux phrases. Pour compléter certains alignements, deux opérateurs sont introduits : *(*s*)*, qui représente 0 ou 1 instance de n'importe quel mot et *(*g*)*, qui représente exactement une instance de n'importe quel mot.

Nous avons appliqué le processus décrit plus haut sur une partie du corpus médical de la campagne d'évaluation des systèmes de Question/Réponse EQueR et extrait ainsi des patrons multi-niveaux pour les quatre relations considérées dans cette étude. Nous donnons à titre illustratif quelques exemples de patrons extraits pour chaque relation :

Maladie – Examen

<examen> en suspicion de <maladie>
<maladie> être *(*g*)* à le <examen>
<examen> pour le NC_GEN *(*g*)* <maladie>
<examen> montre un <maladie>

<examen> *(*g*)* le diagnostic *(*g*)* <maladie>
<maladie> , *(*s*)* <examen>

Maladie – Traitement
<traitement> dans le traitement des <maladie>
<traitement> être *(*g*)* PREP_GENERAL le traitement de le *(*s*)* <maladie>
<traitement> est recommandé pour le traitement des <maladie>
<maladie> , *(*g*)* NC_GEN *(*g*)* une <traitement>
<maladie> nécessitant un <traitement>
<traitement> contre le <maladie>

Maladie – Symptôme
<maladie> , se manifeste par une <symptome>
<symptome> *(*g*)* être *(*s*)* des symptômes d' une <maladie>
<maladie> VERBE_PRINC_INDICATIF *(*s*)* <symptome>
<symptome> *(*s*)* peut VERBE_PRINC_INFINIT la NC_GEN de le *(*s*)* <maladie>
<maladie> *(*g*)* avec <symptome>
<symptom> (<maladie>

Maladie – Médicament
<medicament> est indiqué dans le traitement de la <maladie>
<medicament> , utilisée *(*s*) (*s*)* dans le traitement de <maladie>
<medicament> est un médicament utilisé pour traiter <maladie>
<maladie> chez les NC_GEN traité par <medicament>
<medicament> dans le cas de <maladie>
<medicament> (proposé dans le traitement de *(*s*)* <maladie>

4.2 Extraction et validation des relations sémantiques

Pour acquérir de nouvelles relations sémantiques à partir d'un corpus, *i.e.* de nouveaux couples d'entités liées par une relation identifiée, nous appliquons une démarche en deux temps. Comme dans le cas de l'extraction des patrons de relation, nous commençons par sélectionner des relations candidates en repérant les phrases contenant un couple d'entités intervenant dans une des relations cibles. Dans un second temps, nous confrontons la phrase contenant la relation candidate avec les patrons linguistiques spécifiques de cette relation. Si l'un au moins de ces patrons peut s'appliquer à la phrase considérée, la relation est considérée comme validée. Dans le cas contraire, elle est écartée. Plus formellement, le processus mis en œuvre pour un type de relation est le suivant :

1. appliquer sur le corpus considéré les règles de reconnaissance des entités médicales impliquées dans la relation cible ;
2. extraire du corpus toutes les phrases contenant simultanément les deux entités de la relation cible ;
3. réaliser l'analyse linguistique de chaque phrase sélectionnée, toujours en utilisant l'analyseur LIMA ;

4. remplacer dans chaque phrase les entités par leur type ;
5. pour chaque phrase, calculer sa distance d'édition avec tous les patrons multi-niveaux de la relation. Si la distance d'édition est égale à 0, c'est-à-dire si la relation entre les deux types sémantiques de la phrase respecte le schéma du patron, alors valider la relation.

Nous avons appliqué notre algorithme d'extraction et de validation de relations sémantiques pour les quatre relations retenues dans notre étude sur un corpus de textes médicaux recueillis dans le cadre du projet Technolangue Atonant, corpus différent de celui utilisé pour induire les patrons de caractérisation de ces relations. Voici quelques exemples de relations sémantiques validées par notre méthode (le patron utilisé est introduit par $\Longrightarrow$) :

Maladie – Examen
tomodensitométrie dans le diagnostic des *tumeurs du médiastin*
$\Longrightarrow$ <examen> dans le diagnostic *(*s*)* <maladie>
radiographie pulmonaire pour le diagnostic de *tuberculose*
$\Longrightarrow$ <examen> *(*g*)* le diagnostic *(*g*)* <maladie>

Maladie – Médicament
insuffisance rénale chronique traitée par *Eprex*
$\Longrightarrow$ <maladie> traitée par <medicament>
Le *vaccin* utilisé pour prévenir la *fièvre aphteuse*
$\Longrightarrow$ <medicament> utilisé pour VERBE_PRINC_INFINIT *(*g*)* <maladie>

Maladie – Traitement
chimioprophylaxie contre la *malaria*
$\Longrightarrow$ <traitement> contre la <maladie>
radiothérapie dans le traitement de la *resténose*
$\Longrightarrow$ <traitement> dans le traitement de la <maladie>

Maladie – Symptôme
L'*intoxication* peut provoquer des *vomissements*
$\Longrightarrow$ <maladie> *(*s*)* peut VERBE_PRINC_INFINIT DET_ART_CONTRACT *(*s*)* <symptome>
Botulisme , se manifeste par une *sécheresse de la bouche*.
$\Longrightarrow$ <maladie>, se manifeste *(*s*)* par une <symptome>

5 Évaluation

Dans cette section, nous présentons les résultats des évaluations menées sur deux corpus en français, constitués chacun d'articles scientifiques et de recommandations de bonne pratique médicale téléchargées à partir du site du CISMeF[4]. La première (cf. Table 1) concerne l'identification des entités médicales dans les textes en appliquant les règles de reconnaissance présentées à la Section 3. La seconde (cf. Table 2) porte sur l'extraction et la validation de relations sémantiques grâce à la méthode présentée à la Section 4.2.

La Table 1 résume les résultats obtenus en appliquant nos règles de reconnaissance d'entités

[4]Catalogue et Index des Sites Médicaux Francophones : `http://www.cismef.org`

médicales sur un sous-ensemble d'une taille de 1,5 Mo (soit environ 130.000 mots) du corpus médical de la campagne d'évaluation EQueR. Les mesures utilisées sont classiquement la précision et le rappel, qui se définissent ici de la façon suivante :
– la précision représente le nombre d'entités correctes extraites par notre système sur le nombre total des entités extraites par notre système ;
– le rappel représente le nombre d'entités correctes extraites par notre système sur le nombre total des entités présentes dans le corpus.
La F1-mesure correspond à la moyenne harmonique entre la précision et le rappel. Ces mesures sont réalisées par comparaison avec une annotation manuelle du corpus d'évaluation. Les résul-

Entités sémantiques	Précision	Rappel	F1-mesure
Maladie	0,95	0,80	0,86
Symptôme	0,84	0,76	0,79
Examen	0,94	0,93	0,93
Traitement	0,86	0,81	0,83
Médicament	0,93	0,88	0,90
Moyenne	0,90	0,84	0,86

TAB. 1 – Résultats de la reconnaissance des entités médicales

tats de notre méthode donnés par la Table 1 montrent une précision et un rappel supérieurs ou égaux à 83% en moyenne, ce qui constitue un bon niveau pour ce type de tâche. On peut noter en particulier le niveau élevé de la précision qui caractérise un niveau de fiabilité très significatif. Cette propriété est d'autant plus importante dans le cas présent que la détection des entités sert ensuite de point de départ à l'extraction des relations. Le rappel pourrait quant à lui être amélioré en étant plus exhaustif dans les listes d'entités constituées.

Concernant l'extraction et la validation des relations sémantiques, nous avons appliqué la méthode présentée à la Section 4.2 sur 65 Mo du corpus utilisé dans le cadre du projet Technolangue Atonant, soit environ 10 millions de mots. Les patrons d'extraction appliqués avaient été préalablement appris à partir de la totalité du corpus médical EQueR, soit environ 16 millions de mots. Contrairement au cas des entités, l'annotation manuelle de référence n'a pas été réalisée en parcourant tout le corpus mais en jugeant de la présence effective d'une des quatre relations cibles parmi les phrases abritant des relations candidates, c'est-à-dire les phrases contenant au moins deux entités compatibles avec des relations cibles. Par conséquent, seule la validation des relations candidates est évaluée ici. Pour les mesures d'évaluation, nous avons à nouveau fait appel à la précision et au rappel, définis comme suit :
– la précision représente le nombre de relations validées correctes sur le nombre total des relations validées par notre système ;
– le rappel représente le nombre de relations validées correctes par notre système sur le nombre total de relations annotées dans le corpus.
Comme dans le cas de la reconnaissance des entités, la validation des relations extraites se caractérise par une forte précision et un rappel un peu moins élevé. La différence entre précision et rappel est d'ailleurs plus accentuée dans ce cas que pour la reconnaissance des entités, un peu du fait d'une précision moyenne légèrement plus forte mais surtout par un rappel notablement moins élevé. On peut donc dire que les relations produites par la méthode que nous avons proposée sont globalement d'une bonne fiabilité mais que les patrons linguistiques appris sur le corpus médical EQueR ne couvrent pas toutes les formes par lesquelles les relations cibles

Relations	Précision	Rappel	F1-mesure
Maladie–Examen	0,92	0,63	0,74
Maladie–Médicament	0,91	0,59	0,71
Maladie–Traitement	0,92	0,69	0,78
Maladie–Symptôme	0,90	0,65	0,75
Moyenne	0,91	0,64	0,75

TAB. 2 – Résultats de la validation des relations sémantiques

se manifestent dans le corpus Atonant. La comparaison avec d'autres travaux est quant à elle difficile du fait de la diversité des types de relations considérés, des corpus et des approches adoptées. Néanmoins, il est possible de donner quelques éléments de situation. En utilisant des patrons linguistiques élaborés manuellement pour caractériser des relations d'inhibition dans des phrases extraites de Medline, (Pustejovsky *et al.*, 2002) obtient ainsi une précision de 94% et un rappel de 58,9%. La Table 2 montre que nous obtenons des résultats globalement comparables en construisant ces patrons linguistiques de manière automatique. Le processus de validation des relations extraites peut également être envisagé sous l'angle de la classification : une relation candidate est alors classée comme pertinente ou non pertinente. C'est l'approche retenue par (Craven, 1999) ou par (Rosario & Hearst, 2004). En utilisant un classifieur bayésien naïf sur des relations candidates de type *subcellular–location* extraites de Medline, (Craven, 1999) fait état d'une précision de 78% et d'un rappel de 32%. Dans le cas de (Rosario & Hearst, 2004), le classifieur n'est plus seulement binaire. Il s'agit en effet de discriminer les relations intervenant entre un traitement et une maladie : 8 relations sont ainsi distinguées qui recouvrent la relation *Traite* à laquelle nous nous sommes attachés mais également des relations exprimant qu'un traitement peut prévenir une maladie ou qu'une maladie est un effet secondaire d'un traitement. (Rosario & Hearst, 2004) rapporte les évaluations menées avec plusieurs types de classifieurs et obtient les meilleurs résultats avec un réseau de neurones, la précision étant alors de 96,9%. Il est à noter que ce travail s'appuie sur des ressources plus étendues que le nôtre puisqu'il fait appel à un analyseur syntaxique de surface et qu'il exploite également la ressource sémantique que constitue le MeSH.

6 Discussion

La méthodologie proposée pour l'extraction de relations sémantiques dans le domaine médical repose sur l'identification des entités du domaine puis la validation de relations candidates extraites sur la base de la cooccurrence de ces entités en utilisant des patrons linguistiques. L'utilisation de schémas lexico-syntaxiques pour l'extraction de relations sémantiques a déjà fait l'objet de nombreux travaux. Hearst (Hearst, 1992) est l'une des premières à avoir proposé une approche fondée sur des patrons pour extraire des relations d'hyperonymie. Cependant, sa méthode, qui consiste à extraire un environnement commun à un ensemble de phrases, était essentiellement manuelle. Cette approche a été reprise et complétée par d'autres travaux, toujours dans le domaine de l'extraction de relations sémantiques, dans le but notamment d'automatiser l'extraction des patrons. La méthode développée par Ravichandran (Pantel *et al.*, 2004) dont nous nous sommes inspirés se situe précisément dans cette perspective. Cette démarche s'est également avérée particulièrement productive dans des domaines de spécialité comme en

attestent par exemple les travaux rapportés dans (Finkelstein-Landau & Morin, 1999) ou (Séguéla, 1999), qui se sont focalisés sur des textes techniques.

Bien que se situant dans le droit fil de tous ces travaux, la méthode que nous avons exposée ici s'en différencie par le mode d'application des patrons linguistiques induits. Au lieu de les appliquer à la manière d'expressions régulières, nous calculons une distance entre le patron et la phrase abritant une relation candidate. Cette façon de faire autorise une plus grande souplesse dans l'application des patrons et permet également d'avoir le même mode de fonctionnement lorsque les relations sont caractérisées par des patrons, comme c'est le cas ici, et lorsqu'elles sont caractérisées par des exemples, comme dans une approche de type Memory-Based Learning. On peut même envisager ainsi de mêler les deux approches.

Une autre différence notable avec les travaux tels que (Pantel *et al.*, 2004) est que les résultats de la Section 5 ont été obtenus sans utilisation d'un filtrage *a posteriori* des relations extraites. En dépit de cette absence, la précision se situe à un haut niveau sans que le rappel ne soit trop faible. Plusieurs explications complémentaires peuvent être avancées. Tout d'abord, cette extraction intervient dans un domaine spécialisé et se focalise sur des relations intervenant entre des entités spécifiques à ce domaine. Ensuite, les relations sont de type syntagmatique et non paradigmatique comme dans (Pantel *et al.*, 2004). Enfin, les patrons linguistiques appris restent assez spécialisés puisqu'ils ne sont issus que de la généralisation de couples d'exemples.

7 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode d'extraction de relations sémantiques intervenant entre des entités du domaine médical. Cette méthode utilise des patrons linguistiques multi-niveaux pour valider des relations sémantiques candidates extraites des textes. Ces patrons sont appris automatiquement à partir de textes annotés en s'appuyant sur une notion de distance d'édition étendue.

La méthode proposée ici a montré des résultats encourageants en regard des travaux comparables existants. L'axe principal d'amélioration doit porter sur le rappel. Les évaluations relatives à la validation des relations ont montré que les patrons linguistiques appris ne couvrent pas toutes les manifestations des relations cibles. En outre, étant réalisées seulement à partir des phrases extraites et non de toutes les phrases du corpus d'évaluation du fait de la taille de ce dernier, elles masquent le déficit de rappel résultant de l'absence de reconnaissance des entités médicales déclenchant le processus d'extraction. Même si le niveau de reconnaissance de ces entités peut être considéré comme bon, la nécessité de reconnaître les deux entités d'une relation amplifie l'impact de leur éventuelle mauvaise reconnaissance.

Pour améliorer à la fois la couverture des patrons linguistiques et la reconnaissance des entités médicales, nous envisageons d'adopter une démarche itérative classiquement utilisée dans un tel cas : au lieu de limiter l'usage des patrons linguistiques à la seule validation des relations extraites, il est aussi possible de les utiliser pour extraire de nouvelles entités en ne fixant qu'une seule des entités d'une relation. Ces nouvelles entités viennent à leur tour enrichir la reconnaissance des entités médicales et peuvent ainsi servir à acquérir de nouveaux patrons linguistiques. Une autre voie d'amélioration du rappel est l'utilisation des ressources sémantiques existant dans le domaine médical, comme le thésaurus MeSH ou le méta-thésaurus UMLS. Il

serait ainsi possible d'inclure la vérification de relations sémantiques telles que l'hyperonymie dans la distance d'édition étendue permettant à la fois de construire les patrons linguistiques et de les appliquer. Enfin, parmi les extensions envisagées de ce travail figure également une extension de la couverture des relations de notre ontologie médicale, dont la Figure 1 ne montre qu'une partie. Nous nous sommes limités pour le moment à quatre relations mais les principes testés peuvent tout à fait être appliqués aux autres relations de cette ontologie.

Références

BESANÇON R. & CHALENDAR G. D. (2005). L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASY. In M. JARDINO, Ed., *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, p. 21–24, Dourdan : ATALA LIMSI.

CARABALLO S. A. (1999). Automatic acquisition of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 120–126.

CRAVEN M. (1999). Learning to extract relations from medline. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, Orlando, Florida, USA.

ELY J., OSHEROFF J., EBELL M., BERGUS G., LEVY B., CHAMBLISS M. & EVANS E. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ*, **319**, 358–361.

C. FELLBAUM, Ed. (1998). *WordNet : An Electronic Lexical Database*. The MIT Press.

FINKELSTEIN-LANDAU M. & MORIN E. (1999). Extracting semantic relationships between terms : Supervised vs. unsupervised methods. In *International Workshop on Ontological Engineering on the Global Information Infrastructure*, p. 71–80.

HEARST M. (1992). Automatic aquisition of hyponyms from large text corpora. In *14th International Conference on Computational Linguistics (COLING'92)*, Nantes, France.

MUKHERJEA S. & SAHAY S. (2006). Discovering biomedical relations utilizing the world wide web. In *Pacific Symposium on Biocomputing 11*, p. 164–175.

NÉDELLEC C. (2004). Machine Learning for Information Extraction in Genomics - State of the art and perspectives. In S. SIRMAKESSIS, Ed., *Text Mining and its Applications : Results of the NEMIS Launch Conference*. Springer Verlag.

PANTEL P., RAVICHANDRAN D. & HOVY E. (2004). Towards terascale knowledge acquisition. In *International Conference on Computational Linguistics (COLING'04)*, p. 771–777, Geneva, Switzerland.

PUSTEJOVSKY J., CASTANO J. & ZHANG J. (2002). Robust relational parsing over biomedical literature : Extract inhibit relations. In *PSB 2002*, p. 362–373.

ROSARIO B. & HEARST M. (2004). Classifying semantic relations in bioscience texts. In *42th Annual Conference of the Association for Computational Linguistics (ACL'04)*.

SÉGUÉLA P. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Actes de la conférence Ingénierie des Connaissances (IC'99)*, p. 79–88, Palaiseau.

VINTAR S. & BUITELAAR P. (2003). Semantic relations in concept-based cross-language medical information retrieval. In *ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM)*, Germany.

Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne

Davy WEISSENBACHER, Adeline NAZARENKO
Université Paris-Nord, LIPN, 99 av. J-B. Clément, F-93430 Villetaneuse
{dw,nazarenko}@lipn.univ-paris13.fr

Résumé. On oppose souvent en TAL les systèmes à base de connaissances linguistiques et ceux qui reposent sur des indices de surface. Chaque approche a ses limites et ses avantages. Nous proposons dans cet article une nouvelle approche qui repose sur les réseaux bayésiens et qui permet de combiner au sein d'une même représentation ces deux types d'informations hétérogènes et complémentaires. Nous justifions l'intérêt de notre approche en comparant les performances du réseau bayésien à celles des systèmes de l'état de l'art, sur un problème difficile du TAL, celui de la résolution d'anaphore.

Abstract. In NLP, a traditional distinction opposes linguistically-based systems and knowledge-poor ones, which mainly rely on surface clues. Each approach has its drawbacks and its advantages. In this paper, we propose a new approach based on Bayes Networks that allows to combine both types of information. As a case study, we focus on the anaphora resolution which is known as a difficult NLP problem. We show that our bayesain system performs better than a state-of-the art one for this task.

Mots-clés : réseaux bayésiens, résolution des anaphores, connaissance linguistique, indice de surface.

Keywords: bayesian network, anaphora resolution, linguistic knowledge, surface clue.

1 Introduction

On oppose souvent en TAL les systèmes qui exploitent des connaissances linguistiques et ceux qui reposent sur des indices de surface. Les premiers systèmes ne sont pas toujours fiables parce qu'ils exploitent des connaissances complexes qui peuvent être erronées lorsqu'elles sont calculées automatiquement ou incomplètes lorsqu'elles sont produites manuellement. Les seconds systèmes s'appuient généralement sur des méthodes d'apprentissage automatique et sur des indices de surface qui sont plus faciles à obtenir mais qui ne permettent de traiter que les cas simples ou les plus courants de la tâche dévolue au système.

Dans cet article nous proposons une nouvelle approche qui permet de dépasser cette opposition entre systèmes «pauvres» et système «riches» en connaissances. Cette approche repose sur le formalisme des réseaux bayésiens. Ce formalisme est encore peu exploité en TAL mais il repose sur un modèle probabiliste conçu pour raisonner sur des informations incertaines, partielles et manquantes.

Nous validons notre approche sur la tâche de la résolution automatique des anaphores où, en raison de la complexité et du nombre de connaissances nécessaires, l'opposition des systèmes à base de connaissances linguistiques et d'indices de surface est très marquée. Après avoir validé l'approche en développant un premier classifieur bayésien qui permet de distinguer pronoms impersonnels et pronoms anaphoriques, nous analysons les performances d'un second classifieur qui trouve l'antécédent des pronoms anaphoriques.

La section suivante revient sur les raisons de l'opposition précédente dans le cadre de la résolution des anaphores pronominales. La section 3 décrit le modèle des réseaux bayésiens et son intérêt pour le TAL. Dans la section 4 nous validons notre approche en comparant les performances de différents systèmes pour la distinction des pronoms impersonnels et anaphoriques. Enfin, la dernière section présente un classifieur pour la tâche complète de la résolution des anaphores et compare ses résultats par rapport à l'état de l'art.

2 La complémentarité des connaissances linguistiques et des indices de surface

2.1 Le choix des indices de surface

L'anaphore est une relation linguistique entre deux entités textuelles définie lorsqu'une entité textuelle (l'*anaphore*) renvoie à une autre entité du texte (l'*antécédent*). Comme la présence d'anaphores dégrade considérablement les performances des systèmes de TAL, la question de leur résolution est étudiée depuis longtemps. Ce travail se limite à la résolution de l'anaphore du pronom *it* dans les textes anglais, l'anaphore la mieux connue et la plus facile à résoudre.

L'approche classique pour sa résolution automatique distingue trois étapes : la distinction des pronoms anaphoriques et impersonnels (*it is known that...* vs *it produced...*), la sélection des candidats possibles à l'antécédence et le choix de l'antécédent. Pour chaque étape, les premiers systèmes proposés dans la littérature exploitaient des connaissances linguistiques complexes traduisant les contraintes syntaxiques et sémantiques qui régissent l'anaphore. Comme le calcul automatique de ces connaissances était considéré comme impossible ou trop peu fiable pour être utilisable, ces connaissances linguistiques étaient produites manuellement, ce qui présupposait un important travail d'analyse préalable des textes.

Durant les années 1990, devant le besoin de systèmes de résolution robustes et peu coûteux à mettre en place, un nombre important de systèmes à bases d'indices de surface ont été proposés (Mitkov, 2002). Ces systèmes abandonnent les connaissances linguistiques complexes des premiers systèmes. Ils approchent les connaissances nécessaires par des indices plus simples à calculer et que l'on suppose plus fiables.

Pour la distinction des pronoms anaphoriques, (Husk & Paice, 1987) a ainsi proposé un ensemble d'automates encodant des connaissances linguistiques et permettant de reconnaître les séquences contenant des pronoms impersonnels. Jugeant que ces automates avaient une couverture trop faible, (Evans, 2001) propose une voie alternative reposant sur l'apprentissage automatique des indices de surface pour reconnaître les séquences caractéristiques. Pour le choix de l'antécédent, les connaissances syntaxico-sémantiques sont approchées de la même manière par des méthodes robustes. On sait que les schémas prédicat-argument améliorent les résultats du filtrage (Ponzetto & Strube, 2006), mais comme ces ressources ne sont pas toujours disponibles,

on a cherché à les approcher par un calcul fréquentiel : les régularités des coocurrences entre les sujets, les compléments et les verbes dessinent les contours des classes sémantiques. Les auteurs de (Dagan & Itai, 1990) montrent que les contraintes obtenues peuvent partiellement remplacer les connaissances sémantiques.

2.2 Les limites des indices de surface

Si les indices approchés proposés lors des années 1990 ont permis l'implémentation de systèmes robustes (Mitkov, 2002), leur apport et leurs limites étaient mal connus. Des travaux récents commencent à en mesurer les limites. L'étude de (Kehler *et al.*, 2001) montre ainsi que les fréquences de (Dagan & Itai, 1990) n'améliorent pas les performances d'un système qui exploite déjà des informations morpho-syntaxiques. Les auteurs en concluent que l'apport des fréquences tient davantage du hasard que d'une véritable capture du sens sémantique.

Les limites rencontrées par les systèmes à base d'indices de surface nous renvoient au problème initial. Nous avons besoin de connaissances sémantiques et syntaxiques complexes pour la résolution de l'anaphore pronominale. Ces connaissances linguistiques, lorsqu'elles sont disponibles, ne sont pas fiables. On peut chercher à les remplacer par des indices de surface dont le calcul est toujours réalisable et plus fiable mais ces indices peuvent ne pas exprimer, ou seulement de manière imprécise, les connaissances nécessaires à la résolution, ce qui produit des erreurs.

Nous proposons une modélisation reposant sur les Réseaux Bayésiens (RB), conçu pour raisonner sur des données incertaines et incomplètes. Cette approche probabiliste offre la possibilité d'unifier dans une unique représentation connaissances linguistiques et indices de surface. Cette unification permet de corroborer les connaissances linguistiques grâce aux indices de surface qui sont observés en corpus. A l'inverse, l'exploitation de connaissances linguistiques permet de corriger certaines des erreurs des systèmes à base d'indices de surface.

3 Une approche intégrée : le modèle bayésien

3.1 Des problèmes de classification

La distinction des pronoms impersonnels comme le choix de l'antécédent sont des tâches qui, comme de nombreuses tâches du TAL, se reformulent facilement en problèmes de classification.

Considérons par exemple la classification des pronoms impersonnels et anaphoriques : soit *Corpus* un ensemble de textes d'un même domaine, *Corpus_entraînement* et *Corpus_test* deux sous-ensembles stricts disjoints de *Corpus*, C_1 et C_2 les classes des occurrences des pronoms impersonnels et anaphoriques présents dans *Corpus*. e est une occurrence d'un pronom présent dans *Corpus* décrit par un vecteur $a = v_1, ...v_a$ d'attributs à valeurs dans **R**. Pour les occurrences de *Corpus_entraînement*, les valeurs des attributs v_i sont obtenues à partir d'une analyse humaine du corpus : elles représentent selon les cas des connaissances linguistiques ou des indices de surface.

Le théorème de Bayes dit comment prédire la meilleure classe d'appartenance pour une occurrence d'un pronom inconnu de *Corpus_test* sur la base d'observations faites sur les occurrences

de *Corpus_entrainement*. La classe sélectionnée doit maximiser la probabilité

$$P(C_i|E) = \frac{P(E|C_i)*P(C_i)}{P(E)}$$

où $C_i \in \{C_1, C_2\}$, E une occurrence du corpus de test et $P(C_i|E)$ la probabilité conditionnelle que E appartienne à la classe C_i sachant la valeur des attributs de E, une probabilité estimée à partir des données d'entraînement. Si nous imposons la contrainte d'indépendance des attributs, le classifieur est un «classifieur bayésien naïf». Les attributs étant indépendants, la probabilité $P(E|C_i)$ se décompose en $P(v_1|C_i) * ... * P(v_a|C_i)$ et la probabilité à maximiser se reformule en

$$P(C_i|E) = \frac{P(C_i)}{P(E)} \prod_{j=1}^{a} P(v_j|C_i)$$

Pour tout E de *Corpus_test*, un classifieur bayésien attribue la classe C_1 à l'exemple E si P(Pronom=Impersonnel|E)≥P(Pronom=Anaphorique|E) et la classe C_2 sinon.

3.1.1 Le choix des attributs pour la classification

L'un des premiers systèmes distinguant les pronoms *it* impersonnels et anaphoriques (Husk & Paice, 1987) s'appuie sur un ensemble de règles de logique du 1^{er} ordre pour reconnaître les séquences qui contiennent une occurrence du pronom impersonnel. Les séquences qui introduisent les *it* impersonnels partagent une forme remarquable : elles commencent par un *it* et se terminent par un délimiteur comme *to, that, whether....* Les règles varient selon le délimiteur. Les tests réalisés par Paice montrent que ces règles réalisent un bon score avec 91,4%Acc [1] sur un corpus technique. Cependant les performances sont dégradées si on applique les règles à des corpus de nature différente. Le nombre de faux positifs (FP) augmente : certains attributs sont discriminants sur les corpus techniques mais ne le sont plus sur des corpus de nature différente.

Afin d'éviter cet écueil, (Lappin & Leass, 1994) décrit entièrement les séquences au moyen d'automates à états finis de la forme *It is not/may be<Modaladj>* ; *It is <Cogv-ed> that <Subject>* où *<Modaladj>* et *<Cogv>* dénotent des classes d'adjectifs modaux et de verbes cognitifs connus pour introduire des *it* impersonnels (par exemple *necessary, possible* et *recommend, think*). Ce système a une bonne précision (il produit peu de FP), mais il a un mauvais rappel (il produit beaucoup de FN) : seules les séquences exactes sont reconnues et il est toujours difficile d'obtenir des classes d'adjectifs et de verbes exhaustives.

(Evans, 2001) renonce à exploiter des connaissances linguistiques aussi complexes et se concentre sur des attributs plus fiables, les indices de surface. Evans considère 35 indices syntaxiques et contextuels (ex. la position du pronom dans la phrase, le lemme du verbe suivant...). Un système d'apprentissage, utilisant la méthode des K plus proches voisins, détermine le poids des attributs discriminants pour le domaine du corpus et classe les occurrences inconnues. Les premiers essais réalisent un score de 71,31%Acc satisfaisant sur un corpus de langue générale. (Litran *et al.*, 2004) reproduit un essai identique avec une Machine à Support de Vecteur (SVM) sur un corpus de génomique et obtient un score de 92,71%Acc.

[1] L'exactitude, en anglais Accuracy : Acc= $\frac{VP+VN}{VP+VN+FP+FN}$, où les faux positifs (FP) correspondent aux occurrences d'un pronom anaphorique étiquetées impersonnelles, les faux négatifs (FN) les occurrences de pronoms impersonnels étiquetées anaphoriques, les vrais positifs (VP) et les vrais négatifs (VN) correctement étiquetées comme impersonnels et anaphoriques, respectivement.

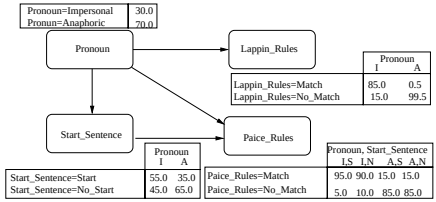


FIG. 1 – Exemple d'un classifieur bayésien modélisé par un réseau bayésien

Ces deux derniers systèmes d'apprentissage reposent donc uniquement sur des indices de surface. Constatant que les connaissances linguistiques sont peu fiables ou incomplètes, les auteurs renoncent à les utiliser comme attributs. Ce choix nous paraît trop radical : dès lors que ces connaissances linguistiques sont pertinentes pour notre tâche, il faut les intégrer dans la décision sous la forme d'attributs mais en se donnant les moyens de raisonner sur des attributs hétérogènes et de qualité variable.

3.1.2 L'inférence sur des attributs imparfaits

Le RB est un modèle conçu pour raisonner sur des attributs incertains et incomplets. Il est composé d'une description qualitative de leurs dépendances, un graphe orienté sans circuits, et d'une description quantitative, un ensemble de probabilités conditionnelles où chaque Variable Aléatoire (VA) est associée à un noeud du graphe. Une 1^{er} étape de paramétrage permet de représenter les connaissances *a priori* pour chaque VA sous la forme d'une table de probabilités conditionnelles. L'étape suivante, l'étape d'inférence, consiste à réviser certaines probabilités *a priori* pour obtenir des probabilités *a posteriori* et à modifier en conséquence les valeurs des VA correspondantes à partir d'observations faites en corpus. Ces nouvelles informations sont propagées au travers du réseau et permettent de réviser les valeurs *a priori* même pour les variables non-observées.

Expliquons sur un exemple très simplifié le mécanisme d'inférence du réseau de la figure 1, un réseau destiné à la classification des pronoms *it*. La 1^{er} étape de paramétrage du réseau, permet de calculer les valeurs *a priori* des probabilités. Sur l'analyse des fréquences d'un corpus d'entraînement ou à partir de l'estimation d'un expert, nous établissons *a priori* qu'environ un tiers des pronoms *it* du corpus sont impersonnels, P(Pronoun=Impersonal)=0,3. Un lien d'influence relie les variables `Pronom` et `Lappin_Rules`, indiquant qu'un *it* a d'autant plus de chance d'être reconnu par une règle de (Lappin & Leass, 1994) qu'il est impersonnel. De même, les liens entre les variables `Pronoun` et `Paice_Rules` d'une part, `Pronoun` et `Start_Sentence` d'autre part indiquent respectivement qu'un *it* a d'autant plus de chance d'être reconnu par une règle de (Husk & Paice, 1987) et d'être en début de phrase qu'il est impersonnel. L'arc (`Start_Sentence, Paice_Rules`) unit les deux variables, car, toujours au regard du corpus d'entraînement ou de l'estimation de l'expert, elles ne sont pas indépendantes. La fiabilité de la règle de (Husk & Paice, 1987) reconnaissant une séquence est augmentée si la séquence est située en début de phrase. Cette influence est mesurée par la table de probabilités conditionnelles associée au noeud `Paice_Rules` de la figure 1.

Une fois l'ensemble des probabilités conditionnelles déterminé, l'étape d'inférence débute.

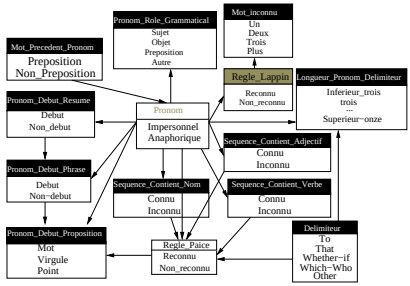


FIG. 2 – Un Réseau Bayésien pour la classification des pronoms *it* impersonnels

Considérons par exemple la phrase *It is well documented that treatment of serum-grown....* Nous appliquons les règles de (Lappin & Leass, 1994) et les règles de (Husk & Paice, 1987) sur cette séquence. Aucune règle de (Lappin & Leass, 1994) ne reconnaît la séquence, nous posons P(Lappin_Rules = No_Match)=1. Une règle de (Husk & Paice, 1987) la reconnaît, nous posons P(Paice_Rules = Match)=1 et comme la séquence se situe en début de phrase nous posons aussi P(Start_Sentence = Start)=1. En représentant graphiquement l'indépendance conditionnelle des VA, le RB permet de compacter la loi jointe globale. A l'aide des probabilités conditionnelles fournies en paramètres nous pouvons inférer la probabilité qui nous intéresse : P(Pronoun=Impersonal|Lappin_Rules=No_Match, Start_Sentence=Start, Paice_Rules=Match)

Du fait qu'une règle de (Husk & Paice, 1987) a reconnu la séquence et que l'occurrence se trouve en début de phrase, le réseau infère une probabilité de 38,9% pour l'occurrence d'être impersonnelle. Nous pouvons modifier cette conclusion en ajoutant d'autres variables au réseau ou en raisonnant avec des observations incertaines ou manquantes. On peut par exemple indiquer que la fiabilité de l'observation est inférieure à 100% et poser P(Lappin_Rules=No_Match)=0,9 pour tenir compte de l'incomplétude des règles de (Lappin & Leass, 1994).

4 1^{re} expérience : l'identification des pronoms impersonnels

4.1 Le protocole expérimental

L'objectif de cette première expérience est de valider notre modèle (on trouvera dans (Weissenbacher & Nazarenko, 2007) une description précise du système développé et une analyse plus complète des résultats obtenus). Nous avons mesuré les performances du Classifieur Bayésien (CB) de la figure 2[2], ainsi que celles du classifieur bayésien naïf (CBN) associé[3], puis nous les avons comparées avec celles des systèmes de l'état de l'art.

[2]Les attributs représentant le fait qu'une règle de (Lappin & Leass, 1994) ait reconnu une séquence sont colorés en gris, en blanc ceux qui correspondent aux règles de (Husk & Paice, 1987), enfin en noir les attributs de (Litran *et al.*, 2004) et (Evans, 2001). Le noeud de prédiction est le noeud `Pronom`, au centre. Il estime la probabilité pour une occurrence donnée de pronom d'être impersonnel ou anaphorique.

[3]Le classifieur bayésien naïf possède les mêmes attributs mais sa structure est différente : le noeud `Pronom` est lié à tous les noeuds et ces derniers ne sont liés à aucun autre.

Méthode	Résultats		
Règles De (Lappin & Leass, 1994)	88,11%	12,8	169,1
Règles De (Husk & Paice, 1987)	88,88%	123,6	24,2
Machine à Vecteurs de Support	92,71%	-	-
Classifieur Bayésien naïf	92,58%	74,1	19,5
Classifieur Bayésien	**95,91%**	**21,0**	**38,2**

TAB. 1 – Résultats des prédictions (Exactitude/Faux Positifs/Faux Négatifs)

Nous avons travaillé sur un corpus de résumé d'articles de génomique construit à partir de la base *Medline* interrogée avec les mots clés *bacillus subtilis, transcription factors, Human, blood cells, gene and fusion*. Nous en avons extrait 11 966 résumés (environ 5 millions de mots) où nous avons identifié 3347 occurrences du pronom *it*. Deux annotateurs humains ont classé chaque occurrence du pronom soit comme anaphorique soit comme impersonnelle. L'accord des annotateurs fut entier après discussion.

Notre corpus étant de taille moyenne, nous avons procédé à une validation croisée pour valider nos résultats. Nous sélectionnons aléatoirement 2/3 du corpus pour calculer les probabilités conditionnelles *a priori*. Nous appliquons ensuite notre CB, ainsi que le CBN, paramétrés grâce à ces probabilités sur le tiers restant. Nous réitérons 20 fois ces opérations pour obtenir une moyenne des performances de chaque système sur le corpus.

4.2 Résultats

Le tableau 1 résume les moyennes des résultats (en exactitude) obtenus par les systèmes de l'état de l'art décrits plus haut[4] et celles des deux classifieurs. Ces résultats montrent que le CB produit une meilleure classification que les autres systèmes, notamment les systèmes à base de règles. Ces résultats valident notre modèle : le CB exploite tous les attributs pertinents et corrige le bruit d'un attribut par la fiabilité des autres. Privé des relations de dépendance entre les attributs, le CBN ne bénéficie pas du mécanisme de correction et surestime leurs fiabilités. Les systèmes à base de règles sont quant à eux entièrement assujettis à la fiabilité des attributs. Les résultats confirment les craintes soulevées dans la section 3.1.1 : on obtient un faible rappel pour les règles de (Lappin & Leass, 1994) et une mauvaise précision pour celles de (Husk & Paice, 1987).

5 2^{nde} expérience : la résolution des anaphores

Assurés des bonnes performances de notre modèle sur la distinction des pronoms impersonnels, nous proposons un classifieur bayésien pour la résolution d'anaphore.

[4]Nous avons ajouté le score du SVM obtenu par (Litran *et al.*, 2004) sur un corpus de génomique similaire pour comparer leurs résultats aux nôtres. Les attributs utilisés par les SVM sont ceux défini par les auteurs. Les valeurs FP et les FN n'ont pas été publiées.

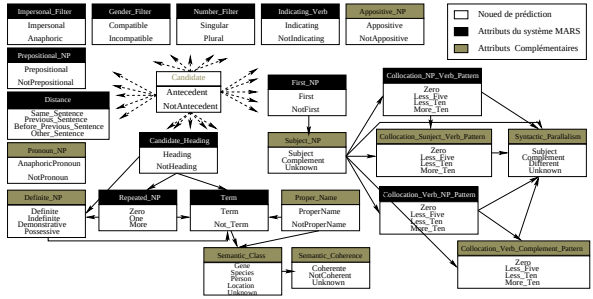


FIG. 3 – Un réseau Bayésien pour la classification des antécédents

5.1 Un classifieur bayésien pour la résolution des anaphores

Nous avons utilisé le système MARS (Mitkov, 2002) comme système de référence pour notre évaluation. Ce système repose sur des indices de surface pour trouver l'élément le plus saillant dans le discours qui précède une occurrence donnée de pronom. Cet élément est celui qui a la plus forte probabilité d'être l'antécédent du pronom. Nous avons ré-implémenté le système en utilisant le même prétraitement des textes que dans notre système bayésien[5] de manière à comparer uniquement les algorithmes des deux systèmes (choix des attributs et mécanisme de prise de décision).

Pour réaliser notre classifieur bayésien (voir figure 3[6]), nous avons conservé tous les indices approchés de MARS (noeuds coloriés en noir sur la figure) mais nous avons ajouté une série d'autres indices (en gris sur la figure) qui sont également pertinents pour le calcul de la saillance et qui sont proposés par plusieurs travaux de l'état de l'art. De notre point de vue, il est en effet utile d'avoir à la fois les indices et les connaissances linguistiques qu'ils approchent. Par exemple, le sujet d'une phrase est souvent l'élément saillant mais comme le calcul du rôle grammatical peut être erroné, il est intéressant d'exploiter en parallèle l'information concernant un indice de surface (*First_NP* : le premier GN de la phrase est très souvent le sujet du verbe) qui peut confirmer ou infirmer l'hypothèse du rôle grammatical.

En suivant un protocole expérimental identique à celui de la section précédente sur le même corpus, nous avons réalisé la résolution avec 4 systèmes différents. Trois systèmes servent de comparaison : le système *Aléatoire* qui choisit un antécédent au hasard dans la liste des candidats, le système *Premier GN* qui sélectionne toujours le premier GN de la phrase précédant le pronom comme antécédent et le système MARS. Le dernier système est le classifieur bayésien (CB) que nous cherchons à évaluer.

Pour les trois derniers systèmes, nous donnons deux mesures différentes des performances, un taux de succès strict et partiel[7]. Le taux de succès est strict lorsque l'antécédent exact a été

[5] Nous avons utilisé dans les deux cas les analyses produites par la plate-forme d'annotation OGMIOS (Derivière *et al.*, 2006).

[6] Le noeud de prédiction est le noeud `Candidat`, au centre. Il estime la probabilité pour une occurrence d'un candidat d'être l'antécédent d'un pronom donné. Ce noeud `Candidate` est lié à tous les noeuds du réseau.

[7] Strict Success rate = $\frac{Anaphorecorrectementrsolue}{Touteslesanaphores}$

Partial Succes rate = $\frac{Anaphorecorrectementetpartiellementrsolue}{Touteslesanaphores}$

annoté par le système et partiel lorsque seule une partie de l'antécédent à été annotée. En raison des erreurs de l'analyse syntaxique en constituants sur laquelle la liste des candidats est calculée, certains GN candidats ne sont identifiés que partiellement ou font défaut. Les performances de nos systèmes ne peuvent atteindre 100%, la dernière colonne donne les scores maximum possibles pour la résolution.

System	**Results**	
	Strict	*Partial*
Aléatoire	6%	-
Premier GN	36.3%	51%
MARS	26.7%	43%
Classifieur Bayésien	44.0%	61%
MAX	93.3%	97.8%

TAB. 2 – Comparaison des résultats (taux de Succès)

La comparaison des scores des systèmes MARS et CB permet d'établir l'apport des connaissances linguistiques complexes dans la résolution en dépit de leur qualité imparfaite. Ces connaissances supplémentaires rendent possible la désambiguisation entre différents candidats. Considérons les phrases *[A grpE heat-shock gene]$_1$ was found by sequencing in [the genome of the methanogenic archaeon Methanosarcina mazei S-6]$_2$. [It]$_1$ is the first example of grpE from the phylogenetic domain Archaea*. Le système MARS attribue des scores identiques pour les candidats 1 et 2 et ne les départage que grâce à l'heuristique du candidat le plus récent, ce qui le conduit à choisir le candidat 2. Le classifieur CB évite cette erreur. La connaissance du sujet et du type sémantique *gène* du candidat 1 augmente à 0.73 sa probabilité d'être l'antécédent du pronom et lève l'ambiguité.

Une analyse détaillée des erreurs du CB montre les limites de notre analyse de la saillance. 47% des erreurs sont dues à un calcul erroné de l'élément saillant : le système ne retrouve pas ce que l'annotateur humain juge «intuitivement» être l'élément saillant parce qu'un nombre plus important d'indices favorisent un candidat différent de l'élément saillant auquel le classifieur associe la plus grande probabilité d'antécédence. Dans 21% des cas, le systèmes trouve bien l'élément qui paraît saillant à l'annotateur humain mais cet élément n'est pas l'antécédent, ce qui met en cause soit notre définition de la saillance soit son rôle dans la résolution de l'anaphore. Dans l'exemple suivant *[Amino acid sequence analysis]$_1$ of [the 33-kDa protein]$_2$ revealed that it is a sigma factor, sigma E*. l'élément le plus saillant est le candidat 1 et il est choisi comme antécédent par le système, une décision qui viole les connaissances du domaine, un facteur sigma est une protéine, des connaissances qu'il faut prendre en compte pour choisir le candidat 2 comme antécédent. Les erreurs restantes proviennent des imperfections des prétraitements linguistiques : principalement des erreurs de segmentation en phrase et de l'analyse syntaxique incorrecte qui ne permet pas de repérer tout les GN candidats.

6 Conclusion

Les réseaux bayésiens présentent un véritable intérêt pour les nombreuses tâches de classification du TAL. Ce modèle permet de dépasser l'opposition historique des systèmes à base de connaissances linguistiques et d'indices de surface. De fait, cette opposition apparaît infondée :

les connaissances linguistiques sont nécessaires mais souvent indisponibles et peu fiables ; les indices de surface sont généralement calculables et de bonne qualité mais il reste des problèmes d'ambiguïté. En unifiant ce deux types de connaissances au sein d'une unique représentation, le modèle offre un mécanisme de raisonnement dont nous nous servons pour corriger et suppléer les connaissances linguistiques en les complétant des indices de surface. Tout l'enjeu consiste selon nous à raisonner sur l'ensemble des connaissances et indices disponibles à un moment donné mais en tenant compte de leur relative fiabilité dans le processus de décision.

Nous avons ensuite validé notre modèle sur le problème de la résolution des anaphores en proposant deux classifieurs, le premier pour distinguer les pronoms impersonnels et anaphoriques, le second pour le choix de l'antécédent. Les résultats de nos classifieurs sont supérieurs à ceux des systèmes de l'état de l'art.

Actuellement seule une expertise linguistique rend compte de la structure des deux classifeurs que nous avons présentés. Nous envisageons de tester les mécanismes permettant d'apprendre la structure même du réseau. Comparer notre structure avec une structure apprise automatiquement devrait permettre de vérifier et d'enrichir la structure du CB actuelle.

Références

DAGAN I. & ITAI A. (1990). Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of COLING'90*, p. 3 :330–332.

DERIVIÈRE J., HAMON T. & NAZARENKO. A. (2006). A scalable and distributed nlp architecture for web document annotation. In *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, p. 56–67.

EVANS R. (2001). Applying machine learning toward an automatic classification of it. *Literary and linguistic computing*, **16**, 45–57.

HUSK G. & PAICE C. (1987). Towards the automatic recognition of anaphoric features in english text : the impersonal pronoun it. *Computer Speech and Language*, **2**, 109–132.

KEHLER A., APPELT D., TAYLOR L. & SIMMA A. (2001). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference*, p. 289–296.

LAPPIN S. & LEASS H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4), 535–561.

LITRAN J. C., SATOU K. & TORISAWA K. (2004). Improving the identification of non-anaphoric it using support vector machines. In *Actes d'International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, p. 58–61.

MITKOV R. (2002). *Anaphora Resolution*. Longman Pub Group.

PONZETTO S. & STRUBE M. (2006). Semantic role labeling for coreference resolution. In *Companion Volume of the Proceedings of EACL'06.*, p. 143–146.

WEISSENBACHER D. & NAZARENKO A. (2007). A bayesian classifier for the recognition of the impersonal occurrences of the *it* pronoun. In *Proceedings of DAARC'07*.

Session
Morphologie

Régler les règles d'analyse morphologique

Bruno CARTONI
ISSCO/TIM/ETI – Université de Genève
40 bd du Pont d'Arve, 1205 Genève
bruno.cartoni@eti.unige.ch

Résumé. Dans cet article, nous présentons différentes contraintes mécaniques et linguistiques applicables à des règles d'analyse des mots inconnus afin d'améliorer la performance d'un analyseur morphologique de l'italien. Pour mesurer l'impact de ces contraintes, nous présentons les résultats d'une évaluation de chaque contrainte qui prend en compte les gains et les pertes qu'elle engendre. Nous discutons ainsi de la nécessaire évaluation de chaque réglage apporté aux règles afin d'en déterminer la pertinence.

Abstract. In this article, we present various constraints, mechanical and linguistic, that can be applied to analysing rules for unknown words in order to improve the performance of a morphological analyser for Italian. To measure the impact of these constraints, we present an evaluation for each constraint, taking into account the gains and losses which they generate. We then discuss the need to evaluate any fine-tuning of these kinds of rules in order to decide whether they are appropriate or not.

Mots-clés : évaluation, analyse morphologique, mots inconnus, morphologie constructionnelle.
Keywords: evaluation, morphological analysis, unknown words, constructional morphology.

1 Introduction

Les mots inconnus représentent un problème récurrent pour le traitement automatique de la langue. Traditionnellement, on distingue trois types de mots inconnus : les noms propres, les erreurs et les mots issus de la créativité lexicale (néologismes), chaque type recevant un traitement particulier. Pour les mots inconnus issus de la créativité lexicale et particulièrement de la créativité morphologique (les néologismes construits), il a souvent été proposé d'employer des règles d'analyse qui formalisent plus ou moins les procédés de construction des mots. Ces règles, souvent basées sur des principes linguistiques, mais également fortement contraintes par les ressources disponibles et d'autres considérations plus mécaniques, rencontrent un certain nombre de limites. En effet, ce type d'approche est rapidement confronté au problème de l'homographie des chaînes de caractères des mots potentiellement construits avec des mots qui ne le sont pas. Ainsi, il arrive que ces règles d'analyse engendrent plus de problèmes qu'elles n'en résolvent. Pour pallier ces écueils, ces

règles sont ajustées au moyen d'un certain nombre de contraintes, linguistiques ou mécaniques.

Dans cet article, nous présentons l'évaluation des contraintes linguistiques et mécaniques qui peuvent être appliquées sur les règles d'analyse des mots inconnus. Nous commençons par passer en revue quelques analyseurs à base de règles qui traitent des mots inconnus morphologiquement construits (section 2), puis nous présentons notre analyseur basé sur les chaînes de caractères et sur des contraintes qui permettent d'éviter les écueils (section 3). Le nombre de mauvaises analyses engendrées par les règles ou par l'application d'une contrainte linguistique peut cependant ternir le bénéfice apporté par les règles dans le traitement des mots inconnus. L'évaluation de notre analyseur et de chacune des contraintes représente par conséquent un sujet crucial. Après avoir posé les jalons de l'évaluation de l'analyse morphologique (section 4.1), nous présentons les résultats de l'évaluation des contraintes de notre analyseur (section 4.2), afin de voir quelles sont celles qui aident le système à tendre vers une performance maximale. Nous concluons en montrant que l'importance donnée aux pertes et aux gains induits par ces contraintes dépend avant tout de la finalité de la tâche.

2 Analyse morphologique et mots inconnus : état de l'art

Les mots inconnus sont un phénomène constant, mais leur proportion relativement restreinte constitue un frein aux méthodes d'apprentissage automatique et nous pousse à privilégier des méthodes davantage linguistiques, à base de règles.

Une grande partie des études qui exploitent les connaissances morphologiques pour traiter les mots inconnus néologiques se concentrent sur l'incomplétude lexicale des lexiques des analyseurs morphosyntaxiques. Elles ont pour principal objectif de deviner la catégorie morphosyntaxique des mots inconnus, le plus souvent en exploitant les terminaisons typiques de ces mots (Guilbaud et al., 1997) et (Woods, 2000). Si certaines études ne se réclament d'aucune approche linguistique particulière, d'autres montrent explicitement leur référence théorique, à l'image de (Byrd, 1983) et (Byrd et al., 1989), qui proposent une application des principes de la morphologie lexématique de (Aronoff, 1976). (Byrd, 1983) prône un véritable calcul morphologique permettant de retrouver la base et le(s) affixe(s) qui constituent le mot construit.

Toutes ces études prévoient d'ajuster les règles par l'entremise de contraintes qui permettent d'éviter les problèmes engendrés par ce type d'approche (cf. ci-après section 3.1). Une des contraintes principales porte sur la catégorie morphosyntaxique de la base, qui peut facilement être filtrée, étant donné qu'elle se trouve dans le lexique. D'autres réglages sont davantage sémantiques, comme l'étymologie latine de certains mots anglais qui favorise l'affixation avec un affixe latin (Byrd et al., 1989).

Du point de vue de l'évaluation, les études mentionnées ne font état ni des gains apportés par ces règles ni de la correction des analyses, et se cantonnent à évaluer d'un point de vue beaucoup plus large l'amélioration des performances de l'analyseur en général. Or, nous pensons que le traitement des mots inconnus est, de par sa nature, une tâche extrêmement précise et granulaire. Une augmentation même minime de la performance de l'analyseur doit être mise en regard non seulement des mauvaises analyses engendrées (le bruit), mais également du nombre de mots non-analysés (le silence). En effet, si une règle d'analyse engendre plus de problèmes qu'elle n'en résout, le gain global est alors insignifiant. Dans la suite, nous présentons les différentes questions soulevées par l'évaluation de l'analyse

morphologique à base de règles. Mais avant cela, nous présentons l'analyseur de mots inconnus préfixés que nous avons contraint puis évalué.

3 Analyseur morphologique basé sur les règles

Dans le cadre d'un projet de recherche plus large de traduction automatique des mots inconnus morphologiquement construits (Cartoni, 2005), nous avons mis au point un analyseur morphologique de l'italien permettant d'analyser les mots inconnus mais corrects construits par préfixation. Par analyse morphologique, nous entendons l'identification des mots réellement préfixés et donc l'individualisation de leur base. Le calcul sémantique base+préfixe est effectué ultérieurement par les règles. Cet analyseur s'appuie sur un lexique de référence de l'italien (*Mmorph* (Petitpierre et al., 1995) – 739 000 formes) dont il extrait les informations lexicales. L'analyseur morphologique est donc constitué de *règles de construction des mots* (RCM[1]), comme le montre la figure 1 :

(1) RCM (iper) :
(2) X = $\text{iper}_{\text{PREF}}$ [Y]
(3) Y ∈ L_{IT}

Figure 1 : Règle d'analyse pour la préfixation en *iper*

La règle de la figure 1 analyse les mots inconnus dont la première séquence de lettres est `i p e r`. Si la séquence de lettres restantes (c'est-à-dire la base potentielle) est présente dans le lexique de référence (ligne (3)), le mot X est alors considéré comme construit. Toutes les recherches que nous avons déjà évoquées (section 2) prenaient évidemment en compte le fait qu'un mot construit l'était avec un affixe (instancié par la règle) et une base qui devait être connue du lexique de référence. Il nous faut toutefois mentionner que cette évidence ne s'applique pas toujours, notamment lorsqu'un préfixe s'accole avec un trait d'union à une base absente du lexique, formant tout de même un mot construit (il s'agit souvent de noms propres, comme *pro-Tibet*). Ajoutons également que, comme les règles traitent des chaînes de caractères, nous devons formaliser plusieurs règles pour un même préfixe en fonction des différents allomorphes qu'il peut avoir (le préfixe *in* peut par exemple prendre la forme *il*, *im*, ou *ir,* suivant la consonne initiale de la base), ou encore avec ou sans trait d'union, l'usage de celui-ci étant passablement flottant.

Une étude de faisabilité (Cartoni, 2006) nous avait déjà montré que, parmi les 46 préfixes productifs de l'italien listés par (Iacobini, 2004), certains sont très peu problématiques en terme de transparence et d'homographie avec d'autres chaînes de caractères. Nous les avons donc pour l'instant mis de côté et nous nous sommes contentés d'implémenter l'analyseur avec les 14 préfixes de l'italien (et leurs allomorphes) qui posent le plus de problèmes, à savoir : *pro, dis, trans, inter, in, poli, arci, retro, post, mini, iper, multi, ri* et *co*.

[1] La morphologie lexématique rejette la notion de mot et préfère parler de lexèmes, comme unité abstraite. Il en résulte que l'on parle maintenant de *règles de construction des lexèmes* (RCL) plutôt que de RCM. Nous conservons cependant l'appellation de RCM, étant donné que, d'un point de vue informatique, il s'agit bien de mots (dans le sens de mot forme) que nous voulons analyser.

3.1 Les problèmes de l'analyse morphologique

De telles règles, bien que très simples, se révèlent relativement efficaces, mais elles présentent évidemment un certain nombre d'écueils, à cause de l'homographie de certaines chaînes de caractères avec des mots réellement construits. (Grabar et al., 2006) distinguent cinq types de mauvaises analyses engendrées par ce genre de méthode : (1) des « lexèmes dans lesquels l'opération étudiée n'est pas la dernière opération constructionnelle », (2) des « lexèmes difficilement analysables comme construits en français », (3) des « lexèmes comportant une suite graphique accidentellement identique aux affixes étudiés », (4) des « lexèmes polysémiques [dont] le sens attesté n'est pas celui qui nous intéresse » et (5) des « erreurs et fautes d'orthographe ».

Dans la présente étude, nous appliquons les règles aux mots inconnus du lexique de référence. Par conséquent, nous rencontrons majoritairement des problèmes du type (1), quand l'opération qui a construit le mot inconnu n'est pas la préfixation (`"prostatiche" = prostata +ico,` et non pas `pro+statiche`), ainsi que des problèmes du type (5) où le mot inconnu est erroné, mais a été analysé comme une séquence préfixe + base (`"progesso"` est la forme erronée de `"progresso"` et non pas `"pro+gesso"`). Il reste néanmoins quelques cas de mauvaises analyses qui proviennent de l'absence du mot dans le lexique de référence bien qu'il ne s'agisse ni d'un néologisme construit ni d'une faute d'orthographe. Il s'agit alors majoritairement d'emprunts ou de termes techniques.

Tout l'enjeu de l'analyseur est par conséquent d'éviter les pièges provoqués par ces homographies. Ainsi, nous avons *réglé nos règles* avec un certain nombre de contraintes que nous décrivons ci-dessous, et que nous évaluons dans la section 4.

3.2 Les contraintes possibles sur les règles

Pour *régler nos règles*, nous avons mis en place deux types de contraintes qui étaient déjà proposées notamment par (Byrd et al., 1986 et Bopp et al., 2004). La première, qui prend en compte la catégorie morphosyntaxique de la base, est basée sur des principes linguistiques. La seconde, qui porte sur la valeur sémantique de la base, est également d'inspiration linguistique, même si certaines données proviennent d'intuitions plus empiriques.

La contrainte de la catégorie morphosyntaxique de la base est motivée par le fait que certains affixes ne s'accolent qu'à certains types de base, même s'il est vrai que plus d'un quart des préfixes de l'italien s'accolent aux trois catégories lexicales majeures (adjectif, nom, verbe) (Iacobini, 2004). Ainsi, pour la règle de préfixation en *mini,* nous avons contraint la règle avec une catégorie de base uniquement nominale. Pour les règles de préfixation en *pro, retro, post, poli, multi, trans, arci,* et *iper*, la base doit être soit adjectivale, soit nominale. Enfin, pour les bases des règles de préfixation en *dis, ri, co,* et *inter,* les trois catégories majeures sont possibles. Notons également que cette contrainte, même si elle est parfois très large, permet d'exclure des mauvaises analyses sur un déterminant ou une conjonction (`arcipel` est un emprunt, et non pas une construction avec le préfixe *arci* et la préposition contractée `pel`).

La deuxième contrainte porte sur la valeur sémantique de la base qui peut favoriser également l'application de tel ou tel préfixe. A moins de disposer de ressources lexicales contenant des informations sémantiques, il est très difficile de formaliser informatiquement ce genre de contrainte. Cependant, la sémantique de la base est parfois observable dans sa forme de

surface. Ainsi, comme l'affirme (Iacobini, 2004, p. 114) « l'emploi d'un préfixe peut être conditionné par le suffixe de la base », étant donné que la valeur sémantique de la base peut être exprimée par l'emploi d'un ou plusieurs suffixes. Cette assertion est intéressante car le ou les suffixe(s) concernés peuvent être exprimés en termes de chaînes de caractères, permettant de contraindre la règle d'analyse. Cette optique revient à dire que certains préfixes sont productifs sur des bases qui sont déjà des mots construits, ou que la « constructivité » des bases permet la préfixation. (Krott et al., 1999) et plus tard (Namer, 2003) ont souligné le nombre important de mots construits sur des bases elles-mêmes construites. Il est donc envisageable, pour certains préfixes, de contraindre la base sur certaines terminaisons que nous considérons alors comme des « indices de constructivité ».

Dans la mesure où certains préfixes comme *ri, co* et *retro* sont réputés productifs avec les noms déverbaux, nous pouvons contraindre les règles d'analyse en imposant la présence de suffixes typiques de la nominalisation déverbalisante, comme *–zione* et *–mento*. Il en va de même pour le préfixe *co* qui est très productif sur des noms d'agent (typiquement suffixés en *–(t)ore*). (Iacobini, 2004) cite également les adjectifs en *-bile* qui sont fréquemment préfixés en *in*. Le corollaire de cette dernière remarque est qu'un grand nombre de noms en *-ità* (suffixation nominale des adjectifs en *-bile*) sont également préfixés en *in* (comme *inconciliabilità, indisponibilità*).

De plus, la préfixation permet la formation d'adjectifs sur des bases nominales, qui prennent la forme de l'adjectif relationnel correspondant. En français, par exemple, *anticancéreux* est formé sur *cancéreux,* qui est l'adjectif relationnel de *cancer*. Même si sémantiquement la base du mot construit est le nom qui est à la base de l'adjectif (*anticancéreux = contre le cancer* et non pas *contre les cancéreux*), la forte régularité de ce type de construction nous permet de considérer la base comme un adjectif relationnel (indépendamment du calcul sémantique nécessaire à l'interprétation du mot construit). Les adjectifs relationnels sont eux aussi des mots construits sur des bases nominales, à l'aide de suffixes typiques de ce genre de formation. Il est donc intéressant de contraindre les bases analysées en fonction des suffixes types de la formation d'adjectifs relationnels. Pour l'italien, les suffixes typiques de formation des adjectifs relationnels sont : *-ale, -are, -ario, -ano, -ico, -ile, -ino, -ivo, -orio, -esco, -asco, -iero, -izio, -aceo* (Wandruszka, 2004). Comme nous travaillons sur des chaînes de caractères, il faut évidemment décupler ces suffixes en fonction de chaque flexion de genre et de nombre (*-ino, -ina, -ini, -ine*).

Ainsi, pour certains préfixes implémentés jusqu'à présent dans notre analyseur, nous avons pu contraindre les règles avec les indices suivants : (a) les indices d'adjectif relationnel pour les préfixes *inter, multi, poli, post,* et *trans ;* (b) les indices de noms d'action (*-zione, -mento*) pour les préfixes *co, retro* et *ri ;* (c) les indices de noms d'agent (*-(t)ore*) pour le préfixe *co,* et enfin, (d) les indices d'adjectifs en *-bile* et de noms en *-ità* pour le préfixe *in.* Evidemment, la validité de ces contraintes doit être vérifiée sur une large échelle, vérification que nous présentons ci-dessous.

4 Évaluation de l'analyse morphologique des mots inconnus

Nous l'avons dit, la plupart des recherches qui exploitent les propriétés morphologiques pour résoudre le problème des mots inconnus construits n'évaluent que le produit final (la couverture lexicale de leur analyseur) ou la vitesse de traitement. Très peu s'intéressent aux gains, et aux erreurs supplémentaires que de telles règles peuvent engendrer. Dans cette

section, nous proposons une *évaluation de progression*, qui permet d'appréhender l'impact de chacune des contraintes appliquées sur les règles.

4.1 Les questions d'évaluation

L'objet de notre évaluation est double. Premièrement, nous voulons évaluer la performance de nos règles avec contraintes, c'est-à-dire le pourcentage d'analyses correctes après l'ajout de chaque contrainte. Idéalement, l'ajout de chaque contrainte devrait augmenter la performance de la règle. Le but ultime de chaque règle est de tendre vers une performance maximale, car une règle qui traite les mots inconnus ne devrait pas fournir d'analyse incorrecte (et générer ainsi plus de bruit que le silence qu'elle réduit). Deuxièmement, nous voulons mesurer plus finement les gains de chaque contrainte (les vrais positifs) par rapport aux nombres de « pertes » provoquées par celle-ci, c'est-à-dire le nombre de mots construits « corrects » mais exclus à cause de l'application de la contrainte (les « faux négatifs »).

Pour évaluer l'impact de ces contraintes sur la performance globale de la règle, nous utilisons comme score minimal (la *baseline*) la performance de la règle contrainte, telle qu'elle est présentée à la figure 1, section 3. Dans l'évaluation plus précise des gains et des pertes, nous cherchons à nous approcher des 100% d'analyse correcte pour les « vrais positifs ».

Pour mener à bien l'évaluation, il faut également décider quelle est la bonne réponse, la réponse attendue. En morphologie constructionnelle, et peut-être encore d'avantage en néologie, il est parfois très difficile de dire si un mot est construit ou non. Comme le soulignent (Schmid et al., 2004) à propos des analyseurs morphologiques de l'allemand, « there is no general agreement yet about what constitutes the correct analyses ». Pour notre part, nous considérons qu'une analyse est correcte quand la base et le préfixe sont trouvés pour un mot réellement construit. La difficulté de décider de la bonne analyse dépend de plusieurs facteurs, et notamment la connaissance approfondie de la règle de préfixation. Cette tâche est d'autant plus complexe que les descriptions théoriques sur les préfixes de l'italien ne prennent pas forcément en considération tous les cas de figure, et que l'italien semble être une langue très flexible morphologiquement (nous y reviendrons).

Pratiquement, pour évaluer notre analyseur et ses contraintes, nous lui avons soumis une liste de mots inconnus de notre lexique de référence qui commencent par les mêmes séquences de lettres que les préfixes étudiés. Ces mots ont été extraits d'un important corpus journalistique de l'italien (Baroni et al., 2004, - environ 380 millions d'occurrences). Les occurrences analysées ont ensuite été réduites en formes uniques. Finalement, chaque forme a été évaluée manuellement pour distinguer les mots construits de ceux qui ne l'étaient pas.

4.2 Évaluation des règles sans contrainte

L'évaluation de la performance de l'ensemble des règles implémentées jusqu'à présent dans notre analyseur permet d'obtenir un score de référence (*baseline*) pour le reste de l'évaluation. Comme le montre le tableau 1, l'analyse par cet ensemble de règles a une performance tout à fait honorable. Mais, en distinguant les formes préfixées avec trait d'union de celles qui ne le sont pas, nous remarquons que les règles sont beaucoup moins performantes quand il n'y a pas de trait d'union.

	mots concernés	analyses correctes	analyses incorrectes
avec trait d'union	2839	2833 (99,79 %)	6 (0,21 %)
sans trait d'union	10191	8962 (87,94 %)	1229 (12,05 %)
total	13030	11795 (90,52 %)	1235 (9,48 %)

Tableau 1 : Évaluation des règles sans contrainte

Notons également que la performance n'est pas uniforme entre les règles et dépend beaucoup du préfixe concerné. Ainsi, les préfixes courts ont plus tendance à se retrouver dans des séquences de lettres ambiguës. Par exemple, la règle pour le préfixe *pro* a une performance de 42 %, alors que la règle de *iper* a une performance de 98,29 %. Pour la suite de l'expérience, nous avons pris en compte uniquement les règles de préfixation sans trait d'union, étant donné leur faible performance.

4.3 Évaluation de la règle avec contrainte de catégorie

Globalement, la mise en place de la contrainte de catégorie sur l'ensemble des règles permet d'améliorer légèrement la performance de toutes les règles. Ainsi, d'une performance globale de 87,94 % d'analyses correctes, nous passons à 89,00 %, même si cette variation est différente dans chaque règle. De plus, le nombre de vrais positifs est acceptable, comme le montre le tableau 2.

	total	vrais positifs	faux positifs
analysés	9955	8898 (89,38 %)	1057 (10,62 %)
	total	faux négatifs	vrais négatifs
pas analysés	238	64 (26,89 %)	174 (73,11 %)

Tableau 2 : Ensemble des règles d'analyse avec contrainte de catégorie

Il nous faut également noter que le nombre de mots réellement préfixés mais exclus par les règles (les faux négatifs) est relativement important. Mais ce phénomène varie également beaucoup selon les règles. Par exemple, pour la règle de préfixation en *inter*, nous avions analysé 505 mots, dont 398 étaient réellement des mots construits (performance globale = 78,81 %). L'application de la contrainte catégorielle sur les trois principales catégories lexicales permet d'exclure des séquences de lettres qui avaient été analysées avec une base n'appartenant pas à l'une de ces trois catégories et qui étaient en fait des mots erronés. Le nombre de mots mal analysés a alors diminué, permettant d'augmenter la performance globale de la règle. Le pourcentage de vrais positifs s'élève alors à 79,5 %, sans pour autant exclure des mots réellement construits (0% de faux négatifs). Dans ce cas, l'application de la contrainte, même si elle n'augmente pas significativement la performance, n'exclut pas non plus de mots corrects.

En revanche, avec le préfixe *multi,* nous avons contraint la catégorie de la base aux seuls noms et adjectifs, comme nous l'indique la présentation linguistique faite par

(Iacobini, 2004). Or, l'application abrupte de ce précepte linguistique provoque l'exclusion de dix formes qui étaient réellement des mots construits. Notre lexique avait analysé leur base comme étant des verbes, alors qu'il s'agissait de participes passés employés comme adjectifs. Il convient par conséquent d'être particulièrement prudent avec l'application de certaines « normes » linguistiques et leur adéquation avec les descriptions linguistiques disponibles.

Cette question de « prescription » théorique se retrouve dans le cas particulier du préfixe *mini,* pour lequel les études morphologiques nous indiquent qu'il ne s'accole qu'à des noms pour former des noms (`RCM(mini) : X/NOM = MINI[Y/NOM]`). Nous avons donc exclu toutes les analyses proposant une base non-nominale. Or, si la contrainte de catégorie permet d'obtenir 98,7 % de vrais positifs, elle exclut de l'analyse 22 formes qui sont en fait des mots construits (des faux négatifs). Si ce silence est dû en partie à des mots qui n'avaient pas une base reconnue comme nom par notre lexique de référence, alors qu'elle aurait due l'être (comme dans *un miniporno*, où *porno* est uniquement enregistré comme adjectif), une autre partie concerne des constructions qui ne sont pas nominales (*minigeografica, miniatomica*). La question est de savoir ici comment considérer ces formations. S'agit-il de nouveaux emplois du préfixe selon un usage qui n'est pas encore enregistré par les études linguistiques ? Cette question renvoie à un plus large débat du TALN actuel qui « s'articule désormais entre les règles postulées et les régularités observées » (Habert et Zweigenbaum, 2002, p. 99).

Nous venons de montrer que l'ajustement des règles d'analyse peut améliorer leur performance, mais que les préceptes linguistiques ne doivent pas forcément être pris en compte de manière aveugle, et qu'une évaluation, même partielle, doit en tous les cas être effectuée à chaque nouveau réglage.

4.4 Évaluation de la contrainte de l'indice de constructivité

Concernant l'indice de constructivité des adjectifs relationnels, la performance globale de la règle diminue fortement - 79,37 % alors que la performance pour les règles sur bases adjectivales avant l'application de la contrainte était de 91,38 %. En revanche, comme le montre le tableau 3, l'application de cette contrainte sur les règles de préfixation étudiées (*multi, poli, post, trans* et *inter*) permet d'obtenir une proportion importante de vrais positifs.

	total	vrais positifs	faux positifs
analysés	680	652 (95,88 %)	28 (4,12 %)
	total	faux négatifs	vrais négatifs
pas analysés	202	154 (76,23 %)	48 (23,77 %)

Tableau 3 : Règles d'analyse avec l'indice de constructivité

Si le pourcentage de vrais positifs augmente vraiment, le pourcentage de faux négatifs est très important (ce qui explique la mauvaise performance globale de la règle) et pourrait remettre en cause l'application d'une telle contrainte. Toutefois, une analyse minutieuse des faux négatifs, nous permet, empiriquement cette fois, d'individualiser un certain nombre de constructions typiques semblant favoriser certaines préfixations (comme les mots suffixés en *–ista* et en *-ismo* très fréquemment préfixés en *pro*). De ce constat linguistique, nous passons

alors à des constats qui relèvent davantage d'intuitions de régularité, mais qui sont sans doute valables et qui peuvent donc être ajoutées aux contraintes. Evidemment, de telles intuitions devraient être évaluées sur un second corpus.

Concernant les autres indices de constructivité (sur les noms déverbaux ou sur les adjectifs en *-bile* et les noms en *-tà*), le pourcentage de vrais positifs est très élevé, comme le résume le tableau 4.

	règle sans indice	indice de constructivité	règle avec indice
in + adj	93,14 %	*-bile*	100 %
in + nom	86,70 %	*-tà*	100 %
co + nom	69,48 %	*-(t)ore* | *-zione* | *-mento*	96 %
retro + nom	90 %	*-zione* | *-mento*	100%
ri + nom	91,21 %	*-zione* | *-mento*	99,65 %

Tableau 4 :Performance des règles avec et sans indice de constructivité

Cependant, le nombre de mots exclus par cette contrainte est, on s'en doute, extrêmement important. En effet, la contrainte exclut à chaque fois plus de la moitié des mots commençant par la séquence de lettres concernée, et parmi eux, entre 50 % et 80 % sont réellement construits, ce qui peut évidemment remettre en cause la contrainte de l'indice de constructivité. Il faut néanmoins souligner la performance quasi-optimale des règles ainsi contraintes et donc l'extrême fiabilité de l'analyse produite pour les vrais positifs.

5 Discussion et conclusion

Nous avons montré que de nombreux moyens linguistiques peuvent être mis en place pour améliorer la performance des règles d'analyse des mots construits. Evidemment, des études plus approfondies, tant linguistiques qu'empiriques, permettraient de découvrir d'autres contraintes plus fines pour régler nos règles. Nous avons également montré que pour certaines règles, les indices de constructivité permettaient d'atteindre une performance maximale.

Si toutes ces contraintes augmentent la performance de l'analyse, nous avons également vu qu'il fallait regarder de plus près toutes les conséquences de l'application de ces contraintes. Parfois, nous avons observé que les études linguistiques ne sont pas toujours des sources pertinentes pour la mise en place de ces contraintes, surtout face à la créativité langagière. Nous avons également montré que si certaines contraintes permettent une amélioration considérable de la performance de la règle, elles excluent aussi beaucoup de bonnes analyses. Il est alors important de resituer les objectifs du système pour décider si la perte est plus dommageable que les gains acquis. Dans cette étude, l'application des contraintes permet d'obtenir une performance maximale (100 % de vrais positifs corrects), ce qui, dans notre projet de traduction automatique, est une condition nécessaire à la poursuite du traitement du mot inconnu. Et les pertes importantes (les faux négatifs) peuvent être ici considérées comme un *statu quo* par rapport à leur condition initiale de mots inconnus.

Références

ARONOFF M. (1976) *Word Formation in Generative Grammar*. Cambridge, The MIT press

BARONI M., BERNARDINI S., COMASTRI F., PICCIONI L, VOLPI A., ASTON G., MAZZOLENI M, (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. Acte de *LREC 2004*, 1771-1774.

BYRD R. J. (1983). Word Formation in Natural Language Processing Systems *IJCAI,* 704-706.

BYRD, R. J., KLAVANS J. L., ARONOFF M., ANSHEN F., (1989). Computer methods for morphological analysis. Actes de *24th ACL*, 120-127.

CARTONI B. (2005). Traduction de règles de construction des mots pour résoudre les problèmes d'incomplétude lexicale en traduction automatique Étude de cas. Actes de *RECITAL 2005*, 565-574.

CARTONI B. (2006). Dealing with unknown words by simple decomposition: feasibility studies with Italian prefixes. Actesde *LREC 2006*, 1674-1677.

GRABAR N., TRIBOUT D, DAL G., FRADIN F., HATHOUT N, LIGNON S., NAMER F., PLANCQ C., YVON F., ZWEIGENBAUM P. (2006). Productivité quantitative des suffixations par -ité et -Able dans un corpus journalistique moderne. Actes de *TALN 2006*, 167-175.

GUILBAUD J.-P., BOITET C. (1997). Comment rendre une morphologie robuste du français encore plus robuste en traitant finement les mots inconnus avec les données disponibles. Actes de *TALN'97*,

HABERT B., ZWEIGENBAUM P. (2002) Régler les règles. *TAL* 43(3) 83-105.

IACOBINI C. (2004). I prefissi. *in La formazione delle parole in italiano*. Grossmann M, Rainer F, (éds). Tübingen, Niemeyer: 99-163.

Krott A., Schreuder R., Baayen R. H. (1999) Complex Words in Complex Words *Linguistics* 37(5), 905-926

NAMER F. (2003). Productivité morphologique, représentativité et complexité de la base: le système moQuête. *Langue française* 140, 79-101

PETITPIERRE D., RUSSEL G, (1995). Mmorph, The Multext Morphology. Genève, Issco (Technical Report).

SCHMID H., FITSCHEN A, HEID U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. Actes de *LREC 2004* 1263-1266

WANDRUSZKA U. (2004). Derivazione aggettivale *in La Formazione delle Parole in Italiano* Grossman M, Rainer F (éds) Tübingen, Niemeyer.

WOODS W. A. (2000). Aggressive morphology for robust lexical coverage. Actes de *Applied natural language processing*

Structures de traits typées et morphologie à partitions

François BARTHÉLEMY[1, 2]
[1] CNAM, Cédric, 292 rue Saint-Martin, 75003 Paris
[2] INRIA, Atoll, 78153 Le Chesnay cedex
barthe@cnam.fr

Résumé. Les structures de traits typées sont une façon abstraite et agréable de représenter une information partielle. Dans cet article, nous montrons comment la combinaison de deux techniques relativement classiques permet de définir une variante de morphologie à deux niveaux intégrant harmonieusement des structures de traits et se compilant en une machine finie. La première de ces techniques est la compilation de structure de traits en expressions régulières, la seconde est la morphologie à partition. Nous illustrons au moyen de deux exemples l'expressivité d'un formalisme qui rapproche les grammaires à deux niveaux des grammaires d'unification.

Abstract. Feature Structures are an abstract and convenient way of representing partial information. In this paper, we show that the combination of two relatively classical techniques makes possible the definition of a variant of two-level morphology which integrates harmoniously feature structures and compiles into finite-state machines. The first technique is the compilation of feature structures into regular expressions, the second one is partition-based morphology. Two examples are given, which show that our formalism is close to unification grammars.

Mots-clés : morphologie à deux niveaux, transducteurs finis à états, structure de traits.

Keywords: two-level morphology, finite-state transducers, feature structures.

1 Introduction

La morphologie à états finis est un courant important de la morphologie informatique qui propose des formalismes de règles contextuelles (grammaires à deux niveaux ou règles de réécriture) pour décrire la morphologie des langues. Ces règles dénotent une relation rationnelle reconnue au moyen d'un transducteur fini.

L'utilisation de structures de traits pour la morphologie à états finis est une pratique relativement courante, que ce soit dans la littérature ou dans les systèmes diffusés comme PC-Kimmo (Antworth, 1995), Xerox Finite-State Tools (Beesley & Karttunen, 2003) ou MMORPH (Petitpierre & Russel, 1995). On peut distinguer deux approches : l'une consiste à compiler les traits statiquement dans les machines finies, l'autre consiste à vérifier les contraintes après exécution de la machine finie au moyen d'une procédure d'unification dynamique. Cette dernière option est coûteuse en temps de calcul à effectuer lors de chaque analyse, mais elle permet d'utiliser

toute la puissance de l'unification. Elle est utilisée notamment dans PC-Kimmo version 2 et dans MMORPH.

La compilation de structure de traits en machines finies impose des contraintes spécifiques qui ont été abordées de deux façons différentes : avec ou sans changement du formalisme des machines finies. Dans la première catégorie, nous trouvons Rémi Zajac (Zajac, 1998) qui propose d'utiliser les structures de traits pour le niveau lexical d'un système de morphologie à deux niveau et remplace sur ce niveau la concaténation par l'unification. Jan Amtrup (Amtrup, 2003) propose quant à lui d'utiliser des machines pondérées par une structure de trait utilisée comme un poids. Cette approche est correcte parce que les structures de traits munies de l'union et de l'unification forment un semi-anneau. La limite de ces deux travaux est que la structure de traits est unique et doit s'enrichir de façon monotone, c'est à dire que les calculs successifs ne peuvent que préciser la valeur des traits, jamais la changer.

L'approche qui consiste à compiler les traits en symboles ordinaires dans des machines finies standards est représentée par XFST d'une part et George Kiraz d'autre part (Kiraz, 1997). Dans XFST, il n'y a pas à proprement parler de structure de traits, mais des traits isolés que l'on peut mentionner à tout endroit dans les expressions régulières pour leur appliquer une opération (fixer, unifier ou redéfinir leur valeur). Ces traits ont une portée globale sur toute une chaîne et les opérations d'évaluation sont effectuées dans un parcours gauche-droite des chaînes. Kiraz propose quant à lui de vraies structures de traits à portée locale, ayant pour seul but un filtrage des règles contextuelle en fonction de traits précisés dans le lexique.

Nous proposons une utilisation plus libre et plus systématique de traits compilés sous forme d'expressions régulières ordinaires, avec la possibilité d'avoir des structures à portée soit locale (par exemple ne concernant qu'un morphème), soit globale (concernant toute une forme), soit encore l'utilisation simultanée de différentes structures de traits ayant des portées différentes. Les traits peuvent être précisés et utilisés aussi bien dans le lexique que dans les règles.

Des restrictions sont apportées à la forme que peuvent prendre les structures de traits ainsi que leurs domaines pour permettre une compilation en expression régulière. Par ailleurs, l'implémentation de la notion de portée d'une structure de trait repose sur les concepts et techniques de la morphologie à partition, une approche de la morphologie à états finis dont le principal contributeur est George Kiraz (Kiraz, 2001).

Dans la section suivante, nous allons voir comment compiler des structures de traits en expressions régulières. Nous verrons ensuite comment ces expressions régulières peuvent être intégrées aux autre composantes d'une description morphologique et nous proposerons un formalisme adéquat. Nous illustrerons l'intérêt de ce formalisme au moyen de deux exemples, l'un n'utilisant qu'une structure de traits globale et l'autre utilisant une véritable grammaire de structure de traits.

2 Compilation des structures de traits

Dans cette section, nous abordons la question de la compilation de structures de traits en automates finis. Plus précisément, nous allons nous intéresser à un sous-ensemble des structures de traits dont la compilation est triviale : il s'agit des structures de traits acycliques prenant leurs valeurs dans des ensembles finis petits.

L'intérêt essentiel de ces structures est d'offrir une syntaxe agréable pour représenter une in-

formation partielle, susceptible d'être complétée via des opérations algébriques (unification ou application de règles).

Dans un premier temps, considérons des structures de traits plates, c'est-à-dire sans structures imbriquées. Chaque trait identifié par son nom prend une valeur dans un ensemble fini de valeurs connu. On peut représenter chaque couple nom-valeur par un symbole spécial et une structure par une chaîne obtenue par concaténation des symboles correspondant à ses différents traits. Prenons par exemple les marques de nombre et personne utiles à décrire la conjugaison du français. Le nombre peut prendre les deux valeurs singulier et pluriel, la personne peut prendre les trois valeurs 1, 2 ou 3. Cela conduit à décrire un alphabet avec les cinq symboles <nombre=singulier>, <nombre=pluriel>, <personne=1>, <personne=2> et <personne=3>. Une structure $[nombre = singulier, personne = 1]$ se compile en la chaîne <nombre=singulier><personne=1>.

Pour assurer l'unicité de la représentation d'une structure, on peut imposer un ordre fixe entre symboles d'une structure basé sur le seul nom des traits, par exemple en utilisant l'ordre lexicographique.

Si l'on connaît à l'avance l'ensemble des traits susceptibles de venir enrichir au fil des calculs un structure de traits, on peut représenter une information partielle au moyen d'une expression régulière représentant l'ensemble des traits. Par exemple, la structure $[personne = 3]$ se compile en (<nombre=singulier>|<nombre=pluriel>)<personne=3>. L'intérêt de cette représentation vient de ce que l'unification de structures de ce genre s'implémente par l'intersection des expressions correspondantes. En définissant une classe de caractères <nom=_> comme l'union des caractères <nom=x> représentants les valeurs que peut prendre le trait `nom`, cette expression peut s'écrire de façon équivalente <nombre=_><personne=3>.

L'unification n'est pas la seule opération que l'on peut désirer réaliser avec des structures de traits. Des règles de grammaires peuvent décrire la construction d'un structure à partir d'une ou plusieurs structures, en spécifiant ce qui doit être emprunté à l'une ou à l'autre au moyen de variables. Par exemple la règle suivante décrit l'adjonction d'un suffixe à une base pourvue de la bonne catégorie syntaxique :

$$\begin{bmatrix}\text{cat} & \boxed{1}\\ \text{nombre} & \boxed{2}\end{bmatrix} \rightarrow \begin{bmatrix}\text{cat} & \boxed{3}\end{bmatrix} \begin{bmatrix}\text{cat} & \boxed{1}\\ \text{de\_cat} & \boxed{3}\\ \text{nombre} & \boxed{2}\end{bmatrix}$$

Le trait `cat` décrit la catégorie syntaxique de la base (premier opérande), du résultat de l'adjonction du suffixe, alors que le trait `de_cat` (second opérande) spécifie la catégorie syntaxique de la base pour que la dérivation soit correcte.

Une telle règle peut être implémentée par un transducteur à trois bandes, une pour chaque opérande et une pour le résultat. Ce transducteur, sous certaines conditions[1], peut être obtenu par intersection de transducteurs implémentant chacun une des variables de la règle. Si l'on suppose que les différents rubans sont synchronisés sur les valeurs des traits, cela donne :
$\boxed{1}$: (_ :_ :_)*(_ :P :P)(_ :_ :_)* where P in <cat=_>
$\boxed{2}$: (_ :_ :_)*(_ :C :C)(_ :_ :_)* where C in <nombre=_>
$\boxed{3}$: (_ :_ :_)*(<cat=X> :_ :_)(_ :_ :_)*(_ :<de_cat=X> :_)(_ :_ :_)*(_ :C :C)(_ :_ :_)*
where X in dom(cat)∩dom(de_cat)

[1] L'intersection de transducteurs n'est pas définie pour les transducteurs en général, mais elle l'est pour certaines sous-classes particulières.

Au sein d'une structure de traits, une variable peut être utilisée pour noter le fait que plusieurs traits partagent une même valeur. Une telle structure est compilée en une disjonction de chaînes, chacune d'elle représentant une des valeurs possibles de la variable.

Le technique de compilation que nous venons de voir s'étend facilement aux structures imbriquées acycliques. Il faut simplement remplacer la notion de nom de trait par celle de chemin. Par exemple, la structure suivante : $\begin{bmatrix} \text{cat} & \text{nom} \\ \text{agr} & \begin{bmatrix} \text{genre} & \text{masc} \\ \text{nombre} & \text{pluriel} \end{bmatrix} \end{bmatrix}$ se compile en <agr.genre=masc><agr.nombre=pluriel><cat=nom>. L'intérêt de cette imbrication est de pouvoir représenter au moyen d'une seule variable l'égalité de tous les traits de la sous-structure.

3 Présentation du formalisme

Le formalisme que nous proposons est basé sur la morphologie à partition. L'historique de ce courant se trouve dans (Kiraz, 2001) alors que sa compilation en automate fini est décrite dans (Barthélemy, 2005). L'idée centrale consiste à définir des relations n-aires dont les différentes chaînes sont divisées en un nombre égal de sous-chaînes. Par exemple, on peut relier une représentation écrite et une représentation phonologique de la façon suivante :

e	x	em	p	l	es
e	gs	ã	p	l	

Comme on le voit, les sous-chaînes mises en correspondances peuvent être de longueurs différentes et éventuellement nulles.

Les relations régulières partitionnées sont la classe de relations qu'on peut décrire avec des expressions régulières augmentées d'une construction nouvelle que nous appellerons tuple, permettant de mettre en relation deux ou plusieurs sous-chaînes. Par exemple, l'expression régulière : $< [lettre]^*, [phoneme]^* >^* < e, \epsilon >$ dénote l'ensemble des chaînes terminées par un e muet. Les opérations comme la concaténation, la disjonction, l'étoile, peuvent intervenir aussi bien à l'intérieur d'un tuple que sur un tuple.

Le formalisme que nous proposons autorise la description de relations n-aires et non seulement binaires, ce qui correspond à une morphologie à n niveaux, n pouvant être différent de 2. Les niveaux supplémentaires peuvent être utilisés soit pour distinguer des facteurs indépendants à un niveau donné, comme c'est le cas par exemple pour la description du Syriaque dans (Kiraz, 2000), soit pour distinguer des niveaux intermédiaires dans une cascade de traitements comme c'est le cas dans l'analyseur morphologique de l'akkadien décrit dans (Barthélemy, 2006).

Le formalisme est fondé sur des expressions régulières étendues pour prendre en compte les notions de partition et de structures de traits. Les règles contextuelles sont admises en tant que raccourcis syntaxiques dénotant des expressions régulières.

Une description comporte les sections suivantes : domaines des traits, types de structures de traits, définition de l'alphabet, types des différents niveaux, types des tuples, autres types d'expressions régulières, définition des machines finies.

Nous allons donner en exemple une description schématique de la conjugaison des verbes français. Nous discuterons cet exemple dans la section suivante. Les points de suspension matérialisent des coupures que nous avons réalisé dans l'exemple pour gagner de la place.

```
PACKAGE verbes;
FEATURES VALUES
   temps: present, futur, passe, imparfait;
   mode: indicatif, subjonctif, conditionnel, imperatif;
   personne: 1, 2, 3;
   nombre: singulier, pluriel;
   conjugaison: 1, 2, 3, irreg;
END VALUES
FEATURE STRUCTURES
   verbe: temps, mode, personne, nombre, conjugaison;
END STRUCTURES
ALPHABET
   [lettre]: a, b, c, d ...
   [voyelle]: a, à, â, e, é ...
   [consonne]: b, c, d ...
END ALPHABET
LEVELS
   1: [verbe:_];
   2: [lettre]+;
   3: [lettre]*;
   4: [lettre]*;
END LEVELS
TUPLES
   <3| LEVEL 3: [lettre], LEVEL 4: [lettre] |3>;
   <2| <3|_|3>* |2>;
   <1| LEVEL 0, LEVEL 1, <2|_|2><2|_|2> |>;
END TUPLES
TYPES
   <radical: LEVEL 1, LEVEL 2, LEVEL 3 > =>
           <1| #1, #2, <2| /LEVEL 3: #3/ |2><2|_|2> |1>;
   <suffixe: LEVEL 2, LEVEL 3 > =>
           <1| _, #1, <2|_|2> <2| /LEVEL 3: #2/ |2>;
END
```

Les domaines de traits sont des listes de valeurs que peuvent prendre les différents traits. D'autres domaines finis de valeurs peuvent également être défini et une même valeur peut appartenir à plusieurs domaines. Les structures de traits sont typées au moyen d'un nom de type associé à la liste des traits de la structure. Dans la syntaxe, le nom de type apparaît en début de structure, suivi de deux points.

Chaque niveau est caractérisé par un numéro et son type est une expression régulière définissant un sur-ensemble des chaînes susceptibles d'être lues sur ce niveau. Le type d'un tuple est constitué de la liste ordonnée de ses niveaux, avec pour chacun d'entre eux une expression régulière restreignant la sous-chaîne pouvant apparaître sur ce niveau. Les différents tuples peuvent différer par leur arité, les niveaux qu'ils comportent et leur degré d'imbrication. D'autres types d'expressions régulières peuvent être définis pour faciliter l'écriture des expressions régulières, par exemple en spécifiant un contenu sous-entendu pour certains niveaux de certains tuples.

Les relations régulières sont nommées. Elles peuvent être définies de trois manières différentes. La première forme est celle d'une expression régulière utilisant des symboles de l'alphabet, des constructeurs de tuples et différentes facilités syntaxiques. Par exemple, on peut utiliser des variables prenant leur valeur dans un ensemble fini. L'expression dénotée est l'union des expressions obtenue par substitution de la variable par une de ses valeurs. Par ailleurs, on autorise l'utilisation d'un joker (wildcard) noté _ dans différents contextes. La projection notée `/LEVEL x: _/` permet de ne spécifier que le contenu d'un niveau dans une expression qui en comporte plusieurs. La construction `REGEXP` réalise implicitement l'union des expressions régulières qu'elle contient, chacune étant terminée par un point-virgule.

```
REGEXP les_radicaux IS
   <radical: lancer, [verbe:conjugaison=1], lanC >;
   <radical: polir, [verbe:conjugaison=2], poli >;
   <radical: pouvoir, [verbe:conjugaison=3], p[OU_EU]v >;
   ...
END
REGEXP terminaisons IS
   <suffixe: [verbe:temps=present,mode=indicatif,
         nombre=singulier,personne=1|3,conjugaison=1], e>;
   <suffixe: [verbe:temps=present,mode=indicatif,
         nombre=singulier,personne=1|2,conjugaison=2|3], s>;
...
END
LET formes=intersect(les_radicaux,terminaisons);
```

Le deuxième moyen de spécifier une relation régulière est par application d'opérations sur des relations définies auparavant. Les opérations comprennent les opérations ensemblistes (union, intersection, différence) et les opérations rationnelles (concaténation, étoile). La projection permet d'éliminer certains niveaux. Sous certaines conditions, l'opération de jointure permet de composer deux relations ayant des domaines différents.

Le dernier moyen de décrire une relation régulière est l'utilisation de règles contextuelles. Ce sont des adaptations aux relations n-aires des règles classiques de la morphologie à deux niveau. Les règles de coercition spécifient un certain motif et contraignent les valeurs que peuvent prendre, en contexte, les sous-chaînes filtrées par ce motif. Par exemple, la règle suivante décrit la réalisation d'un méta-caractère C susceptible de s'écrire ç ou c selon le contexte (comme par exemple dans le verbe lancer, je lançais) :

```
 <3| C, _ |3> => <3| C, ç |3>
    IF _ <2| <3_|3>* XXX |2><2| <3|_, [lettre]-(i|e)|3> _ |2> _
```

Le motif apparaît à gauche de la flèche et la restriction à droite de la flèche. Le contexte est décrit en utilisant `XXX` pour désigner le centre de la règle, à distinguer de _, utilisé ici comme joker. Dans notre système multi-niveaux, il n'y a pas de distinction explicite entre niveau lexical et niveau de surface. N'importe quel ensemble de niveaux peut être précisé dans le motif et deux règles différentes peuvent utiliser des ensembles de niveaux différents, ce qui introduit plus de souplesse et justifie le changement de nom de coercition de surface en coercition tout court.

Une règle de restriction de contexte décrit un contexte dans lequel un motif peut exclusivement apparaître (syntaxe : `motif ONLY IF contexte`). Une règle composite est un règle qui cumule les deux contraintes de coercition et de restriction de contexte.

L'utilisation de règles contextuelles posent des problèmes de conflits, quand deux règles sont d'une certaine façon contradictoires. Nous ne traiterons pas de ce problème en détail dans cet article dont ce n'est pas l'objet. L'existence de ces conflits justifie qu'on considère les règles comme un ensemble et non séparément. La détection des conflits peut être automatisée (Beesley & Karttunen, 2003) et leur résolution peut être aidée par une procédure interactive.

4 Exemple avec structure de traits unique

Dans ce premier exemple, nous voulons insister sur la question du niveau abstrait d'une représentation et promouvoir l'idée que la multiplicité des niveaux permet d'offrir une réponse adéquate. Ce qu'on appelle le niveau lexical dans un système de morphologie à deux niveaux traditionnel est une représentation relativement concrète sur laquelle il faut appliquer quelques transformations pour obtenir une représentation de surface. Il s'agit en fait d'une approximation aussi précise qu'on peut faire de la représentation de surface d'un morphème avant application des mécanismes de dérivation et/ou de flexion.

Dans un système comme PC-Kimmo, la représentation abstraite de la forme est ce qu'on appelle la glose (gloss), une chaîne de caractère précisée dans le lexique et destinée à être affichée en réponse à certaines requêtes. Nous proposons d'inclure cette information dans un ou plusieurs niveaux, sans exclure les deux représentations classiques : approximation avant composition et forme de surface.

Rémi Zajac (Zajac, 1998) propose d'utiliser une structure de traits comme niveau abstrait d'un système à deux niveaux. Nous allons affiner cette idée pour permettre une compilation en machine finie : il faut représenter sous forme d'une structure de trait uniquement les traits élémentaires prenant leur valeur dans un ensemble fini et petit et sous la forme d'une chaîne utilisant un niveau spécifique les informations structurées ou ayant un grand nombre de valeurs.

Dans l'exemple de la conjugaison du verbe français, une forme abstraite doit préciser le lemme, le temps, le mode, la personne, le nombre. Le nombre de lemmes est grand. Créer un symbole par lemme conduit à multiplier le nombre de symboles au-delà de ce qui est couramment accepté par les implémentations de machines finies. Le lemme sera donc noté sous la forme d'une chaîne de caractère et cela constitue le niveau 1. Les autres informations ont peu de valeurs, on les regroupe donc dans une structure de traits qui occupe le niveau 2. Les niveaux 3 et 4 sont consacrés aux représentations intermédiaire et de surface. Par ailleurs, pour coordonner les lemmes et les terminaisons, il faut connaître le paradigme de conjugaison utilisé. Cette information pourrait être mise sur un niveau de service, mais pour simplifier la description, nous la plaçons dans la structure de traits du niveau 1.

Cet exemple illustre comment le typage permet de ne préciser que l'information pertinente pour les radicaux et les terminaisons, tout en ayant une représentation sous-jacente unique, ce qui permet d'opérer une intersection. Cette intersection réalise l'unification des structures de traits spécifiées dans les deux expressions régulières, et notamment l'identification de leur unique trait commun, `conjugaison`. Par exemple,
`<radical: lancer, [verbe:conjugaison=1], lanC >` est une notation équivalente à l'expression :

```
<1| lancer, [verbe:conjugaison=1],
      <2| <3|l,_|3><3|a,_|3><3|n,_|3><3|C,_|3> |2><2|_|2> |1>.
```

Il convient ensuite de compléter la description en précisant comment relier le niveau intermédiaire (niveau 3) avec la réalisation de surface (niveau 4).

```
RULE SET
  <3| $L,$L |3> where $L in [lettre];
  <3| C, c |3>
     ONLY IF _ XXX /Level 3: (e|i) _/;
  <3| s,_ |3> => <3| s,x |3>
     IF <1| [verbe:temps=present,nombre=singulier,
             mode=indicatif],
             pouvoir|vouloir, <2|_|2><2| XXX |2> |1>;
  ...
```

La dernière règle illustre comment une règle contextuelle peut être conditionnée par la valeur des traits en utilisant simplement la notion de contexte habituelle.

5 Exemple avec plusieurs structures de traits

Nous allons prendre comme exemple une grammaire ayant une structure linéaire, décrivant une morphologie basée exclusivement sur des suffixes. Les machines finies permettent de représenter plus facilement de telles structures que des arbres quelconques. Une morphologie basée à la fois sur des préfixes et des suffixes, voire des circonfixes, est plus difficile à traiter. Ces problèmes techniques ne sont pas insurmontables, mais alourdiraient trop notre exemple. Nous allons donc nous limiter à des suffixes susceptibles de changer la catégorie syntaxique d'un mot et donc son type de flexion.

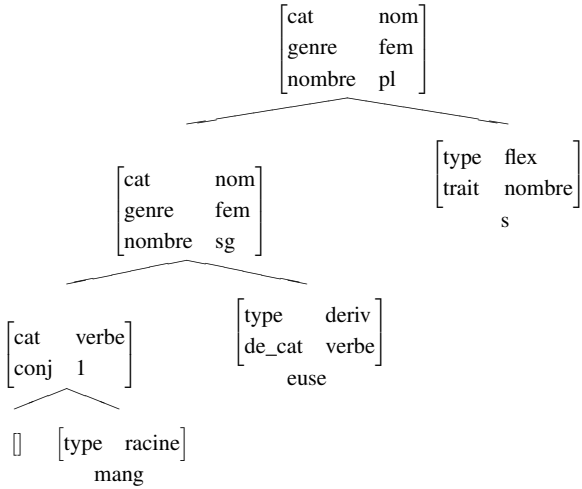


FIG. 1 – Exemple de structure

La figure 1 donne un exemple de structure que nous voulons représenter, celle qui correspond à la forme `mangeuses`. Dans cette structure binaire, des informations doivent être propagées comme par exemple le genre, qui n'est pas modifié par la marque de pluriel. Les suffixes de flexion modifient une partie de la structure, préservant le reste, alors que les suffixes de dérivation bloquent la transmission d'information qui n'est plus pertinente compte tenu du changement de catégorie syntaxique. C'est ici le cas du paradigme de conjugaison.

Nous voyons dans cet exemple que le nombre de noeuds internes de la structure est égal au nombre de morphèmes, ce qui permet d'utiliser le même tuple pour décrire un morphème et son père. Dans chacun de ces tuples, il y aura donc deux structure de traits : une associée au seul morphème, l'autre concernant la structure regroupant le morphème et tous ceux qui le précèdent.

```
REGEXP affixes IS
<affixe: [verbe:cat=verbe,conj=1], [racine], mang >;
<affixe: [nom:cat=nom,genre=fem,nombre=sg],
                              [deriv:de_cat=verbe], euse >;
<affixe: [nom:cat=nom,nombre=pl], [flex:trait=nombre], s >;
...
END
RULE SET composition IS
  <affixe: _, [deriv:de_cat=$A], _ >
      ONLY IF _ <affixe: [_:cat=$A], _ , _> XXX _;
  <affixe: [nom:genre=$G], [flex:trait=nombre], _ >
      ONLY IF _ <affixe: [nom:genre=$G],_,_> XXX _;
  ...
```

La dérivation peut être traitée au moyen d'une règle contextuelle unique qui vérifie que la base possède la catégorie syntaxique requise. Pour la flexion, en revanche, il faut une règle pour chaque type de suffixe, car les traits propagés et les traits révisés ne sont pas les mêmes.

Les deux règles règles données en exemple illustrent comment les différentes structures de traits interagissent et notamment comment certains traits en sont unifiés. Elles pourraient aussi bien s'exprimer sous forme de règle de grammaires de traits. Par exemple la seconde :

$$\begin{bmatrix} \text{nom} & \\ \text{genre} & \boxed{1} \\ \text{nombre} & \boxed{2} \end{bmatrix} \rightarrow \begin{bmatrix} \text{nom} & \\ \text{genre} & \boxed{1} \end{bmatrix} \begin{bmatrix} \text{flex} & \\ \text{nombre} & \boxed{2} \\ \text{trait} & \text{nombre} \end{bmatrix}$$

6 Conclusion

Dans cet article, nous montrons comment l'utilisation simultanée de deux techniques préexistantes, à savoir la compilation de structure de traits en chaînes de caractères et la morphologie à partition, offre un pouvoir de description intéressant.

Il n'y a bien sûr aucune augmentation de puissance du formalisme. Il s'agit de facilité d'écriture : les structures de traits sont pratiques parce qu'on ne précise que l'information connue et que l'ordre des traits n'est pas significatif. De plus ce formalisme est familier aux personnes travaillant dans le TAL.

La technique que nous proposons offre un risque d'explosion de la taille des machines. Ce risque est important si l'on multiplie les traits, les valeurs et surtout les unifications entre structures éloignées. Notre expérience montre que ce risque n'est pas rédhibitoire. Nous avons réalisé un prototype qui compile une description syntaxique en un automate fini, en utilisant la boîte à outils FSM (Mohri *et al.*, 2002). Nous avons écrit des grammaires relativement grosses ($\sim$ 50 règles) sans provoquer d'explosion incontrôlée (Barthélemy, 2006).

L'outil de Xerox (xfst) offre une possibilité intéressante pour éviter l'explosion combinatoire : elle consiste à choisir entre un calcul statique ou dynamique pour les valeurs de traits. Dans le cas d'un calcul dynamique, ce ne sont pas les seules valeurs de traits qui sont représentées sous forme de symboles dans les machines, mais les calculs à réaliser sur ces traits lors d'une évaluation de gauche à droite. A priori, il semble possible d'adapter cette technique à notre formalisme.

Notre proposition permet une utilisation plus générale des traits que les travaux antérieurs proposant une compilation en machine finie. Par rapport à (Zajac, 1998), (Amtrup, 2003), l'apport principal est la notion de portée d'une structure qui peut être locale à un tuple, ce qui autorise la multiplicité de structures ayant certains traits communs dont les valeurs sont indépendantes. Les interactions entre structures de traits sont plus riche que dans (Kiraz, 1997). Par rapport aux approches qui proposent une évaluation dynamique des structures de traits, les gains proviennent d'une meilleure intégration avec les calculs d'automates (par exemple, calcul d'intersection) ainsi qu'une plus grande efficacité.

Références

AMTRUP J. W. (2003). Feature structures as weights in finite state morphology. In *FSMNLP*, Budapest, Hongrie.

ANTWORTH E. L. (1995). User's guide to pc-kimmo version 2.

BARTHÉLEMY F. (2005). Partitioning multitape transducers. In *International Workshop on Finite State Methods in Natural Language Processing (FSMNLP)*, Helsinki, Finlande.

BARTHÉLEMY F. (2006). Un analyseur morphologique utilisant la jointure. In *In Traitement Automatique de la Langue Naturelle (TALN'06)*, Leuven, Belgique.

BEESLEY K. R. & KARTTUNEN L. (2003). *Finite State Morphology*. CSLI Publications.

KIRAZ G. A. (1997). Compiling regular formalisms with rule features into finite-state automata. In *ACL*, Madrid, Espagne.

KIRAZ G. A. (2000). Multitiered nonlinear morphology using multitape finite automata : a case study on syriac and arabic. *Computational Linguistics*, **26**(1), 77–105.

KIRAZ G. A. (2001). *Computational Nonlinear Morphology*. Cambridge University Press.

MOHRI M., PEREIRA F. C. N. & RILEY M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, **16**(1), 69–88.

PETITPIERRE D. & RUSSEL G. (1995). Mmorph : the multex morphology program.

ZAJAC R. (1998). Feature structures, unification and finite-state transducers. In *FSMNLP'98*, Ankara, Turquie.

Analyse morphosémantique des composés savants : transposition du français à l'anglais

Louise DELÉGER[1], Fiammetta NAMER[2], Pierre ZWEIGENBAUM[3, 4]
[1] INSERM, UMR_S 872, Éq. 20, Les Cordeliers, 75006 Paris
Université Pierre et Marie Curie-Paris6, UMR_S 872, 75006 Paris
Université Paris Descartes, UMR_S 872, 75006 Paris
[2] ATILF et Université Nancy 2, CLSH, 54015 Nancy
[3] CNRS, UPR3251, LIMSI, 91403 Orsay
[4] INALCO, CRIM, 75343 Paris Cedex 07
louise.deleger@spim.jussieu.fr,
fiammetta.namer@univ-nancy2.fr, pz@limsi.fr

Résumé. La plupart des vocabulaires spécialisés comprennent une part importante de lexèmes morphologiquement complexes, construits à partir de racines grecques et latines, qu'on appelle « composés savants ». Une analyse morphosémantique permet de décomposer et de donner des définitions à ces lexèmes, et semble pouvoir être appliquée de façon similaire aux composés de plusieurs langues. Cet article présente l'adaptation d'un analyseur morphosémantique, initialement dédié au français (DériF), à l'analyse de composés savants médicaux anglais, illustrant ainsi la similarité de structure de ces composés dans des langues européennes proches. Nous exposons les principes de cette transposition et ses performances. L'analyseur a été testé sur un ensemble de 1299 lexèmes extraits de la terminologie médicale WHO-ART : 859 ont pu être décomposés et définis, dont 675 avec succès. Outre une simple transposition d'une langue à l'autre, la méthode montre la potentialité d'un système multilingue.

Abstract. Medical language, as many technical languages, is rich with morphologically complex words, many of which take their roots in Greek and Latin – in which case they are called neoclassical compounds. Morphosemantic analysis can help generate decompositions and definitions of such words, and is likely to be similarly applicable to compounds from different languages. This paper reports work on the adaptation of a morphosemantic analyzer dedicated to French (DériF) to analyze English medical neoclassical compounds, and shows the similarity in structure of compounds from related European languages. It presents the principles of this transposition and its current performance. The analyzer was tested on a set of 1,299 compounds extracted from the WHO-ART terminology: 859 could be decomposed and defined, 675 of which successfully. Aside from simple transposition from one language to another, the method also emphasizes the potentiality for a multilingual system.

Mots-clés : analyse morphosémantique, composition savante, terminologie médicale.

Keywords: morphosemantic analysis, neo-classical compounding, medical terminology.

1 Introduction

La plupart des vocabulaires spécialisés, et en particulier le vocabulaire médical, comprennent une part importante de lexèmes morphologiquement complexes, construits à partir de racines grecques et latines, qu'on appelle « composés savants ». Segmenter ces composés en lexèmes de base est la tâche de l'analyse morphologique. Lorsque celle-ci inclut à la fois une partie formelle et une partie sémantique, on parle d'analyse morphosémantique. Ce type d'analyse est particulièrement adapté aux composés savants, où le sens est souvent « compositionnel », c'est-à-dire qu'il est la combinaison au moins partielle du sens des composants du lexème complexe. L'analyse morphosémantique est donc utile pour les méthodes intéressées par la sémantique, comme la génération de définitions ou la détection de termes similaires.

Il a de plus été observé que la structure morphologique des lexèmes composés savants est similaire dans de nombreuses langues européennes (Iacobini, 2003). Il semble donc possible d'appliquer une analyse linguistique dédiée aux composés savants d'une langue à d'autres langues proches. (Namer, 2005a) l'a montré pour un certain type de composés médicaux en proposant une analyse des noms de pathologies (comme HYPERCALCIURIE) pouvant être appliquée au français, à l'allemand, à l'espagnol, à l'italien et à l'anglais. L'analyse morphosémantique de tels composés montre ainsi un potentiel multilingue.

Dans le domaine médical, plusieurs travaux se sont intéressés à l'analyse de ces lexèmes complexes. Les premiers se concentrent sur un type particulier de règles de formation des lexèmes, comme les règles de suffixation en -ITIS (Pacak *et al.*, 1980) ou -OSIS (Dujols *et al.*, 1991), puis élargissent leur champs d'analyse (Wolff, 1984). (Lovis *et al.*, 1995) décomposent les termes médicaux en introduisant la notion de morphosémantèmes, unités ne pouvant être décomposées sans perdre leur sens original. On trouve une notion similaire dans le système Morphosaurus (Schulz *et al.*, 1999; Markó *et al.*, 2005). Cet outil, qui ne se limite pas aux composés savants, est l'un des rares fonctionnant en multilingue ; il ne va cependant pas jusqu'à l'interprétation sémantique. (Iavindrasana *et al.*, 2006) utilisent un outil statistique de segmentation morphologique (Creutz *et al.*, 2005) pour mettre en correspondance les termes similaires d'une terminologie médicale (WHO-ART). L'outil DériF (Namer & Zweigenbaum, 2004) effectue une analyse morphosémantique des lexèmes dérivés ou composés français et produit une décomposition hiérarchique (par opposition à (Markó *et al.*, 2005) ou (Lovis *et al.*, 1995) où la segmentation reste linéaire) ainsi qu'une définition sémantique des lexèmes et un ensemble de lexèmes sémantiquement apparentés (relations d'hyponymie ou d'équivalence, par exemple). Son potentiel pour une application multilingue a été souligné dans (Namer, 2005b).

Cet article présente nos travaux concernant l'adaptation de DériF aux composés savants médicaux anglais[1]. Notre but est de transposer l'analyse par DériF des composés français à l'anglais, afin d'illustrer la similarité des mécanismes des composés savants dans des langues européennes proches et d'obtenir un outil qui fait défaut sur l'anglais, que nous baptisons DériA. Ce travail peut en outre être vu comme une première étape vers un système multilingue.

Nous posons dans un premier temps les éléments théoriques sous-jacents à ce travail, puis décrivons l'analyseur morphosémantique et notre liste de lexèmes test. Nous expliquons ensuite les modifications effectuées sur le système et son mode d'évaluation. Nous exposons nos résultats et discutons la méthode, puis concluons avec quelques perspectives.

[1] Cet article actualise une précédente version de nos travaux soumise à la conférence MEDINFO 2007.

2 Eléments théoriques

La base de ce travail est l'analyse morphosémantique, c'est-à-dire une analyse morphologique associée à une interprétation sémantique. Nous cherchons à lier le lexème d'entrée à sa base (en cas de dérivation) ou à ses composants (en cas de composition). La décomposition est associée à une description du sens du lexème complexe basée sur le sens de ses composants. Un lexème complexe est formé à partir de n'importe quelle combinaison des deux règles suivantes :

– la **dérivation** qui se manifeste formellement par l'ajout d'un affixe (préfixe ou suffixe) à une base ; par exemple : $\text{BOSSE}_{\text{NOM}}$ / $\text{BOSSU}_{\text{ADJ}}$; $\text{GRIPPE}_{\text{NOM}}$ / $\text{ANTIGRIPPE}_{\text{ADJ}}$
– la **composition** consiste à former un lexème en combinant deux composants, qui peuvent être chacun soit des lexèmes de la langue moderne, soit des racines grecques et latines appelées éléments de formation (EF) ; par exemple THERMORÉGULATION, ARTHRALGIE.

Nous avons choisi ici de traiter les composés savants anglais (formés à partir d'EF). Cependant, comme nous l'avons expliqué, un lexème complexe peut avoir été formé à la fois par composition et par dérivation. Un composé peut donc être la base d'une règle de dérivation ; de même, un lexème dérivé peut être l'un des composants d'une règle de composition. C'est pourquoi nous devons prendre en compte les lexèmes composés et dérivés (comme HAEMORRHAGIC).

Notre hypothèse de travail pour la transposition de l'analyse du français vers l'anglais est que l'on peut appliquer une même analyse linguistique aux composés savants de plusieurs langues. En effet, nous supposons que ces composés sont formés de la même façon et que les composants mis en jeu sont similaires, les principales différences étant orthographiques (par exemple, ALGIE en français et ALGIA en anglais). Les difficultés éventuelles se situeraient :

– au niveau de l'ordre combinatoire des composants : l'analyse ne fonctionnera pas si l'ordre n'est pas le même dans les deux langues. Ce cas devrait être peu fréquent, car il s'agit ici de composition classique qui reprend l'ordre latin ou grec ;
– au niveau des composants eux-mêmes : la correspondance des analyses ne peut se faire que si les lexèmes anglais et français sont bien dans les deux cas formés à partir d'EF. C'est notre hypothèse et l'analyse pourra se faire car ces EF sont listés et en nombre limité.
– au niveau de la jointure des composants : la modification du premier composant lors de sa combinaison avec le deuxième (allomorphie), par exemple le rajout de l'élément de jointure *-o-*. Si ces phénomènes sont très différents, il peut y avoir des problèmes dans l'analyse. Nous supposons qu'ils sont similaires, mis à part quelques modifications orthographiques ;
– au niveau des processus morphologiques de suffixation et préfixation appliqués aux composés savants. En effet, les affixes ne sont pas les mêmes dans les deux langues. Cependant, nous supposons que, dans le cas des composés savants, il est suffisant de remplacer les affixes français par des affixes anglais de même « classe » (par exemple, les suffixes formateurs d'adjectifs relationnels français peuvent être remplacés par leurs homologues anglais).

3 Matériel de départ

3.1 L'analyseur morphosémantique DériF

Nous nous basons sur la version française de l'analyseur morphosémantique DériF (« Dérivation en Français »), destiné aussi bien à la langue générale qu'à des vocabulaires plus spécialisés

arthralgia/NOM	cardiomegaly/NOM	crystalluria/NOM
atelectasis/NOM	cerebellar/ADJ	dermatomyositis/NOM
blepharospasm/NOM	claustrophobia/NOM	dextrocardia/NOM
calcinosis/NOM	clostridial/ADJ	dorsal/ADJ
capillary/ADJ	cryptococcal/ADJ	dysmenorrhea/NOM

TAB. 1 – Extrait de la liste de lexèmes construits contenant un élément de formation

tels que la langue médicale. L'analyse effectuée est purement linguistique et implémente un certain nombre de règles de décomposition ainsi que de schémas d'interprétation sémantique des lexèmes. Les ressources nécessaires à l'outil comprennent des lexiques de lexèmes lemmatisés et étiquetés morphosyntaxiquement et une table des éléments de formation. Lorsque DériF est appliqué à un vocabulaire biomédical, le système ne se limite pas à une simple décomposition et interprétation sémantique, mais produit également un ensemble de lexèmes lexicalement reliés.

Le système prend en entrée une liste de lexèmes étiquetés et lemmatisés et produit en sortie :

1. une représentation ordonnée des règles d'analyse qui se sont appliquées successivement ;
2. une définition (ou glose) du lexème en langue naturelle ;
3. une catégorie conceptuelle, inspirée des descripteurs principaux du thésaurus MeSH (anatomie, organisme, maladie, etc.) ;
4. un ensemble de lexèmes qui sont potentiellement liés lexicalement au lexème d'entrée. Les différentes relations identifiées sont l'équivalence (*eql*), l'hyponymie (*isa*) et la relation de proximité sémantique « voir-aussi » (*see*).

Un exemple d'analyse morphosémantique pour le lexème ACRODYNIE est donné ci-dessous (*N* signifie *nom* et *N** est associé à un EF nominal).

ACRODYNIE/N ==> (1) [[acr N*] [odyn N*] ie N]
(2) douleur (de–lié(e) à) articulation
(3) maladie
(4) eql : acr/algie, eql : apex/algie, see : acr/ite, see : apex/ite

3.2 Le corpus de test : la terminologie WHO-ART

Pour tester DériA, notre transposition de DériF à l'anglais, une liste de lexèmes a été sélectionnée. Ceux-ci ont été extraits de la terminologie WHO-ART, qui décrit les effets secondaires des médicaments, car une des applications visées par ce travail est d'apporter une contribution au domaine de la pharmacovigilance. Nous avons segmenté les termes anglais en lexèmes, car DériF traite des lexèmes simples et non des unités polylexicales. Parmi ces lexèmes, nous n'avons retenu que les composés savants (nous cherchons en effet à adapter DériF pour ce type de lexèmes). Cette sélection a été effectuée à la fois automatiquement, en rejetant tous les lexèmes de 4 caractères ou moins (ils ne sont quasiment jamais construits), et manuellement en parcourant la liste pour extraire les composés savants. Nous avons ainsi obtenu une liste de 1299 lexèmes sur un total de 3476. Comme expliqué précédemment, nous avons sélectionné à la fois les composés « purs » et les lexèmes dérivés contenant un EF. Un extrait de la liste est donné dans le tableau 1. Afin d'avoir une idée de la proportion des différents types de composés, nous avons calculé leur pourcentage sur un échantillon de 200 lexèmes. Nous avons 45 % de composés purs et 55 % de lexèmes dérivés.

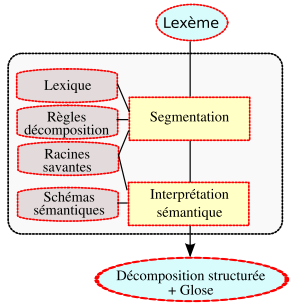


FIG. 1 – Analyse morphosémantique

Les lexèmes ont été préalablement étiquetés morphosyntaxiquement et lemmatisés avec l'étiqueteur Treetagger[2]. Un lexique de lexèmes étiquetés, extrait du Specialist Lexicon de l'UMLS[3] a été utilisé pour aider Treetagger à traiter les lexèmes inconnus.

4 Transposition de DériF en anglais

4.1 Modifications effectuées

Le mécanisme de l'analyse morphosémantique de DériF est schématisé sur la figure 1. Nous avons identifié plusieurs niveaux où l'analyse fait appel à des éléments spécifiques à la langue :

1. *Le lexique* de lexèmes étiquetés qui est utilisé par le système pour tester l'existence d'un composant et pour obtenir sa catégorie grammaticale.
2. *La table des EF*, où chaque élément est associé à un lexème en français moderne décrivant son sens, à un type conceptuel, à une catégorie grammaticale et à un ensemble d'éléments liés par des relations d'équivalence, d'hyponymie ou de proximité sémantique.
3. *Les règles de décomposition*, qui sont déclenchées dans un certain ordre selon la catégorie grammaticale du lexème et l'affixe identifié (s'il y en a un). Ces règles identifient la tête et les autres composants, en se basant sur la table des EF et le lexique. Chaque règle inclut des exceptions et des normalisations orthographiques sur les composants.
4. *Les schémas d'interprétation sémantique*, qui produisent des gloses à partir de la relation entre la tête du lexème et ses autres composants.

Nous avons apporté des modifications pour chacun de ces niveaux :

1. Nous avons remplacé le lexique par un lexique anglais du Specialist Lexicon de l'UMLS.
2. Nous partons de l'hypothèse que les EF sont les mêmes en français et en anglais, avec seulement de légères différences orthographiques. Nous avons donc effectué de petites

[2] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`
[3] `http://www.nlm.nih.gov/pubs/factsheets/umlslex.html`

EF	Anglais	Type	Relations lexicales
algia	pain	disease	odyn, algo, ~itis
blephar	eyelid	anatomy	palpebr, <-ocul, ~coro
ectomy	surgical excision	act	~tomy, ~stomy
gastr	stomach	anatomy	stomac, gaster, ~hepat, ~entero, <-abdomino, ~pancreat

TAB. 2 – Table des éléments de formation (extrait)

modifications sur les EF français (par exemple, retirer les accents, comme pour BLÉPHARO / BLEPHARO) pour obtenir les équivalents anglais. Les types conceptuels attribués à chaque EF auraient pu être conservés en français, mais nous avons décidé que les traduire en anglais serait plus adapté, ce qui a été fait aisément car ces types sont peu nombreux. Nous avons également associé des lexèmes en anglais moderne à chaque EF, que nous avons obtenus à partir de deux listes de racines savantes, une extraite du Specialist Lexicon de l'UMLS, l'autre du dictionnaire médical Dorland[4], les racines étant appariées avec leur équivalent en anglais moderne dans ces deux listes. Nous avons mis en correspondance automatiquement les EF et validé les résultats. Les EF pour lesquels aucun équivalent n'a pu être obtenu ont été traités manuellement. Enfin, l'ensemble des EF lexicalement liés ont été remplacés par leurs équivalents anglais (suivant les modifications orthographiques effectuées pour obtenir les EF anglais). Les catégories grammaticales ont été conservées. Un extrait de cette table est fourni dans le tableau 2. Les relations lexicales entre les éléments sont étiquetées comme suit : <- pour une relation d'hyponymie, ~ pour une relation voir-aussi, et aucun signe pour une relation d'équivalence.

3. Nous sommes très peu intervenus au niveau des règles de décomposition puisque notre hypothèse est que les composés savants sont formés de la même manière. Nous avons donc simplement adapté les exceptions, les normalisations orthographiques et les affixes (par exemple, le suffixe français -IQUE est remplacé par le suffixe anglais -IC).
4. Enfin, nous avons traduit les schémas d'interprétation sémantique afin que ceux-ci génèrent des gloses anglaises. Par exemple, le schéma suivant est associé au suffixe -IA (-IE en français) et au préfixe HYPER- avec les composants Y/X comme noms : *Affection of **X** linked to the excess of **Y*** (où X et Y sont remplacés par les équivalents modernes des EF ou par des lexèmes simples du lexique, comme par exemple HYPERCALCEMIA dont l'analyse produit *« Affection of blood linked to the excess of calcium »).*

4.2 Évaluation

Nous avons testé DériA sur les 1299 lexèmes extraits de la terminologie WHO-ART. La sortie attendue pour chaque lexème est donc une décomposition hiérarchique des lexèmes, une définition en langage naturel, un type conceptuel et un ensemble de lexèmes liés. Nous avons évalué la couverture, la précision et le rappel de ces résultats. Nous avons défini la couverture comme la proportion de lexèmes que le système analyse, la précision comme le rapport entre le nombre d'analyses correctes et le nombre total d'analyses, et le rappel comme le rapport entre le nombre d'analyses correctes et le nombre d'analyses attendues (i.e. le nombre de lexèmes). Une analyse

[4] http://www.merckmedicus.com/pp/us/hcp/thcp_dorlands_content.jsp?pg=/ppdocs/us/common/dorlands/dorland/dmd-a-b-000.htm

est considérée valide si sa décomposition et sa définition sont correctes ; nous n'avons pas pris en compte les lexèmes lexicalement liés ni le type conceptuel dans cette évaluation.

La méthode de (Iavindrasana *et al.*, 2006) est appliquée, comme la nôtre, à la terminologie WHO-ART, de sorte que nous mis en place une comparaison. Nous avons soumis la même liste de lexèmes à Morfessor (qui est l'outil utilisé dans ces travaux) dans les mêmes conditions que (Iavindrasana *et al.*, 2006) et avons évalué couverture, précision et rappel.

5 Résultats

Au stade actuel de DériA (DériF en anglais), nous avons pu obtenir 859 analyses sur les 1299 lexèmes de notre liste (voir tableau 3), ce qui donne une couverture de 66 %. Des exemples d'analyses sont donnés dans le tableau 4. On observe que le système a pu analyser aussi bien de purs composés savants (*[[arthr N*] [algia N*] N]*) que des lexèmes dérivés construits à partir d'une base savante (*[a+ [dactyl N*] +y N]*).

Nous avons identifié plusieurs causes pour lesquelles un lexème n'a pas pu être décomposé :

– Certaines règles de suffixation ne sont pas actuellement implémentées dans DériF et DériA : c'est le cas pour les suffixes -ATION et -ISM, les lexèmes tels que LACRIMATION et HERMAPHRODITISM ne sont donc pas décomposés ;
– Certains éléments ne sont pas présents dans la table des EF ni dans le lexique de lexèmes : par exemple CAMPT- n'est pas dans la table des EF et le lexème CAMPTODACTYLY n'a donc pas été décomposé ;
– Des erreurs sont survenues pendant le pré-traitement (c'est-à-dire des lexèmes mals étiquetés) : par exemple, CORPORAL a été étiqueté comme nom au lieu d'adjectif.

Nous avons mesuré une précision de 78,5 % (voir tableau 3) ce qui est assez satisfaisant. Etant donnée une couverture modérée, cela donne un rappel de 52 %. Les erreurs sont dues à :

– Une mauvaise structuration de la décomposition. On en voit un exemple dans le tableau 4 avec le lexème MENINGOENCEPHALITIS, dont la bonne décomposition devrait être comme suit : [**[** [mening N*] [encephal N*] **]** [itis N*] N]. L'élément -ITIS devrait être la tête de la combinaison des deux EF MENING- et ENCEPHAL-, ce qui donnerait une définition telle que *« inflammation related to head and meninges. »*
– Une définition insatisfaisante, ce qui est souvent dû au fait que le sens du lexème n'est pas suffisamment reflété par celui de ses parties. Le lexème ACANTHOSIS (voir tableau 4) en est une illustration : sa décomposition est correcte mais son sens s'est opacifié et ne peut être dérivé de celui de ses composants. Il devrait être de nos jours analysé comme non construit.
– Une erreur d'étiquetage : les lexèmes mal étiquetés n'ont pas pu être correctement analysés. C'est le cas de ALVEOLAR (dernière ligne du tableau 4) qui a été traité comme un nom dérivé d'un adjectif (ALVEOLAR en tant qu'adjectif est, lui, correctement analysé par le système).

Aucun cas de bruit ni de silence ne semble être dû à une spécificité des composés anglais par

Nb total de lexèmes	Lexèmes décomposés (couverture)	Nombre d'analyses correctes	Précision	Rappel
1299	859 (66 %)	675	78,5 %	52 %

TAB. 3 – Résultats de l'évaluation de DériA

Lexème/Cat	Décomposition	Définition	Type	Lexèmes proches
arthralgia/N	[[arthr N*] [algia N*] N]	pain (of–linked to) joint	disease	eql : arthr/algesia see : arthr/itis
adactyly/N	[a+ [dactyl N*] +y N]	Affection characterized by the absence of digit	disease	
gastroeso-phageal/ADJ	[[gastr N*] [oesophag N*] al ADJ]	Related to oesophagus and stomach	anatomy	eql : stomac/oesophag isa : abdo-min/oesophag
meningoence-phalitis/ADJ	[[mening N*] [[encephal N*] [itis N*] N] N]	(Part of – Specific type of) encephalitis related to meninges		
acanthosis/N	[[acanth N*] [osis N*] N]	(Part of – Specific type of) disease related to prickle	disease	
alveolar/N	[[alveolar A] N]	Entity being alveolar		

TAB. 4 – Exemples d'analyses de lexèmes effectuées par DériA (correctes puis incorrectes)

rapport aux composés français.

L'évaluation des résultats de l'analyse des lexèmes par Morfessor (méthode de (Iavindrasana *et al.*, 2006)) donne une couverture de 93,7 %, une précision de 53,2 % et un rappel de 49,9 % (voir tableau 5). Nous avons également calculé la précision des deux outils sur leur intersection de couverture, soit 830 lexèmes, et avons obtenu 58,7% pour Morfessor et 78,2% pour DériA.

6 Discussion

La précision de notre système adapté (DériA) est encourageante, et pourrait être améliorée en identifiant les mots composés dont le sens s'est figé. Le rappel est plus bas mais devrait augmen-

Nb total de lexèmes	Lexèmes décomposés (couverture)	Nombre d'analyses correctes	Précision	Rappel
1299	1217 (93,7 %)	648	53,2 %	49,9 %

TAB. 5 – Résultats de l'évaluation de Morfessor

ter rapidement en implémentant des règles supplémentaires. Les résultats sont similaires à ceux obtenus sur le français par (Namer & Baud, 2006). Le système ne génère pas plus d'analyses erronées que dans la langue d'origine. De plus, l'analyse des résultats (section 5) n'a pas relevé d'erreurs dues à une différence de mécanisme dans la formation des composés. Les problèmes potentiels énumérés (section 2) ne semblent pas s'être réalisés. Bien que notre corpus de test ne soit pas exhaustif, nous en concluons que ces difficultés sont rares et pourraient, le cas échéant, être traitées grâce à des règles d'exception. Ce système spécifique à une langue a donc pu être adapté avec succès à une autre langue pour l'analyse de composés médicaux.

L'utilisation d'un analyseur morphosémantique comme DériF basé sur des méthodes linguistiques présente en outre un certain nombre d'avantages. Le système effectue à la fois une décomposition morphologique et une interprétation sémantique, tandis que d'autres méthodes restent au niveau de la segmentation morphologique ou ajoutent de la sémantique après l'application d'un outil de décomposition. Nous produisons également une décomposition hiérarchique et non simplement linéaire. Les résultats obtenus en suivant la méthode de (Iavindrasana *et al.*, 2006) ont donné une couverture bien plus grande qu'avec DériA, mais une précision très inférieure (20-25 % de moins) et un rappel légèrement moindre. Ceci montre qu'un nombre presque identique de lexèmes sont correctement analysés (rappel), mais que DériA est bien plus exact car il propose beaucoup moins d'analyses incorrectes. En se basant sur un analyseur statistique, la méthode de (Iavindrasana *et al.*, 2006) a l'avantage d'être indépendante de la langue ; cependant, l'implémentation utilisée fait également usage d'une table de morphèmes, de sorte que la transposition vers une autre langue n'est pas immédiate non plus. De plus, la segmentation sur des bases statistiques cause également certains types d'erreurs qui peuvent être facilement évitées avec des règles linguistiques comme celles de DériA (par exemple, CHEMOSIS est segmenté en C+HEM+OSIS par Morfessor). De plus, comme nous l'avons souligné ci-dessus, DériA propose en sortie une information plus riche que Morfessor (simple segmentation linéaire).

Ce travail offre aussi la perspective d'un système capable de fonctionner avec plusieurs langues. Nous avons transposé le système du français à l'anglais, mais une prochaine étape pourrait être de faire fonctionner le système sur les deux langues (utiliser le même système aussi bien pour analyser des composés anglais que français) ou de l'utiliser pour la traduction : produire une définition anglaise d'un lexème français (et vice-versa). Il faudrait pour cela une table multilingue de racines (comme proposé dans (Namer, 2005b) et préparé ici pour l'anglais et le français), des schémas d'interprétation sémantique multilingues (comme obtenus ici pour ces deux langues) et une traduction des lexèmes du lexique. Un tel système pourrait, par exemple, contribuer à la recherche d'information translingue, avec les mêmes principes que dans (Markó *et al.*, 2005).

7 Conclusion

Dans cet article, nous avons montré comment transposer un analyseur morphosémantique basé sur des règles linguistiques depuis le français vers l'anglais afin d'analyser des composés savants médicaux. Ceci vérifie l'hypothèse que les composés savants de différentes langues peuvent être analysés de la même façon sur le cas du voculaire médical de deux langues de type romane (le français) et germanique (l'anglais). La méthode peut être appliquée sur les composés savants d'autres vocabulaires spécialisés et devrait donner des résultats similaires. Ce travail constitue également un premier pas vers la création d'un système multilingue, qu'on obtiendrait en appliquant la méthode à d'autres langues, tâche plus ou moins aisée selon la proximité des langues (le français et l'anglais étant proches, on peut supposer que de nouvelles difficultés se poseront

avec d'autres langues plus éloignées).

D'un point de vue directement applicatif, nous disposons désormais d'un système fonctionnant sur l'anglais que nous pouvons utiliser dans le domaine de la pharmacovigilance sur les termes de la terminologie WHO-ART afin de mesurer leur proximité, ce qui constituerait une alternative à la méthode proposée par (Iavindrasana *et al.*, 2006).

Références

CREUTZ M., LAGUS K., LINDEN K. & VIRPIOJA S. (2005). *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Rapport interne, Computer and Information Science, Helsinki University of Technology.

DUJOLS P., AUBAS P., BAYLON C. & GRÉMY F. (1991). Morphosemantic analysis and translation of medical compound terms. *Methods of Information in Medicine*, **30**, 30–35.

IACOBINI C. (2003). Composizione con elementi neoclassici. In M. GROSSMAN & F. RAINER, Eds., *La formazione delle parole in italiano*, p. 69–96. Tübingen : Niemeyer.

IAVINDRASANA J., BOUSQUET C. & JAULENT M.-C. (2006). Knowledge acquisition for computation of semantic distance between WHO-ART terms. In *Stud Health Technol Inform*, p. 839–44.

LOVIS C., MICHEL P., BAUD R. & SCHERRER J. (1995). Word segmentation processing : a way to exponentially extend medical dictionaries. *Methods of Information in Medicine*, p. 28–32.

MARKÓ K., SCHULZ S. & HAHN U. (2005). Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, p. 537–545.

NAMER F. (2005a). Guessing the meaning of neoclassical compounds within LG : the case of pathology nouns. In *Proceedings of Generative Approaches to the Lexicon*, p. 175–84, Geneva.

NAMER F. (2005b). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. In M. JARDINO, Ed., *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, p. 63–72, Dourdan : ATALA LIMSI.

NAMER F. & BAUD R. (2006). Defining and relating biomedical terms : towards a cross-language morphosemantics-based system. *Int J Med Inform*.

NAMER F. & ZWEIGENBAUM P. (2004). Acquiring meaning for French medical terminology : contribution of morphosemantics. In M. FIESCHI, E. COIERA & Y.-C. J. LI, Eds., *MEDINFO*, p. 535–539, San Francisco.

PACAK M., NORTON L. & DUNHAM G. (1980). Morphosemantic analysis of -itis forms in medical language. *Methods of Information in Medicine*, **19**, 99–105.

SCHULZ S., ROMACKER M., FRANZ P., ZAISS A., KLAR R. & HAHN U. (1999). Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Proceedings of MIE'99*, Ljubliana, Slovenia : IOS Press.

WOLFF S. (1984). The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. *Methods of Information in Medicine*, **4**(23), 195–203.

Session
Traduction

A tool for detecting French-English cognates and false friends

Oana FRUNZA, Diana INKPEN
School of Information Technology and Engineering
University of Ottawa, Ottawa, ON, K1N 6N5, Canada
{ofrunza,diana}@site.uottawa.ca

Résumé. Les congénères sont des mots qui ont au moins un sens en commun entre deux langues en plus d'avoir une orthographie semblable. La reconnaissance de ce type de mots permet aux apprenants de langue seconde ou étrangère d'enrichir plus rapidement leur vocabulaire et d'améliorer leur compréhension écrite. Toutefois, les faux amis sont des paires de mots qui à l'écrit ont des similarités, mais ils ont des significations différentes. Pour leur part, les congénères partiels sont des mots qui ont la même signification dans certains contextes dans chacune des deux langues. Cet article présente une méthode pour la classification automatique des paires des mots classées en congénères ou faux amis, en utilisant des mesures de similarité orthographiques et des méthodes d'apprentissage automatique. Ainsi, nous construisons des listes complètes des congénères et des faux amis entre les deux langues. Nous désambiguisons les congénères partiels dans des contextes spécifiques. Nos méthodes sont évaluées pour le français et l'anglais, mais elles seraient applicables à d'autres paires des langues. Nous avons construit un outil qui prend ces listes et marque dans un texte français les mots qui ont des congénères ou des faux amis en anglais, dans le but d'aider les apprenants en français langue seconde ou étrangère à améliorer leur compréhension écrite et à développer une meilleure rétention.

Abstract. Cognates are pairs of words in different languages similar in spelling and meaning. They can help a second-language learner on the tasks of vocabulary expansion and reading comprehension. False friends are pairs of words that have similar spelling but different meanings. Partial cognates are pairs of words in two languages that have the same meaning in some, but not all contexts. In this article we present a method to automatically classify a pair of words as cognates or false friends, by using several measures of orthographic similarity as features for classification. We use this method to create complete lists of cognates and false friends between two languages. We also disambiguate partial cognates in context. We applied all our methods to French and English, but they can be applied to other pairs of languages as well. We built a tool that takes the produced lists and annotates a French text with equivalent English cognates or false friends, in order to help second-language learners improve their reading comprehension skills and retention rate.

Mots-clés : congénères, faux amis, congénères partiels, mesures de similarité orthographiques, apprentissage automatique, apprentissage des langues assisté par ordinateur.

Keywords: cognates, false friends, partial cognates, orthographic similarity measures, machine learning (ML), computer-assisted language learning (CALL).

1 Introduction

When learning a second language, a student can benefit from knowledge in his/her first language (Gass, 1987), (LeBlanc *et al.*, 1989). Cognate words can accelerate vocabulary acquisition and facilitate the reading comprehension task. On the other hand, a student has to pay attention to pair of words that are false friends and partial cognates. The following definitions are language-independent, but the examples that we give are for French and English, the focusses of our work.

Cognates, or True Friends (Vrais Amis), are pairs of words that are perceived as similar and are mutual translations. The spelling can be identical or not, e.g., *nature* - *nature*, *reconnaissance* - *recognition*. Some researchers refer to cognates as being pairs of words that are orthographically identical and to near-cognates as the ones that have slightly different spelling. In our work, we adopt the cognate definition for both.

False Friends (Faux Amis) are pairs of words in two languages that are perceived as similar but have different meanings, e.g., *main* (= *hand*) - *main* (meaning *principal* or *essential*), *blesser* (= *to injure*) - *bless* (that is translated as *bénir* in French).

Partial Cognates are pairs of words that have the same meaning in both languages in some but not all contexts. They behave as cognates or as false friends, depending on the sense that is used in each context. For example, in French, *facteur* means not only *factor*, but also *mailman*, while *étiquette* can also mean label or sticker, in addition to the cognate sense.

Although French and English belong to different branches of the Indo-European family of languages, their vocabularies share a great number of similarities. Most of these similar words penetrated the French and English language due to the geographical, historical, and cultural contact between the two countries over many centuries — and here we talk about borrowings. Most of the borrowings have changed their orthography, following different orthographic rules (LeBlanc & Séguin, 1996) and most likely their meaning as well.

Cognates have been employed in Natural Language Processing (NLP) for different tasks. The applications include sentence alignment (Simard *et al.*, 1992), (Melamed, 1999), and improving statistical machine translation models (Marcu *et al.*, 2003). Machine Translation (MT) systems can benefit from extra information when translating a certain word in context. Knowing if a French word is a cognate or a false friend with an English word can improve the translation results. Cross-Language Information Retrieval systems can use the knowledge of the sense of certain words in a query in order to retrieve desired documents in the target language.

We focus on the automatic identification of cognates and false friends. Our approach is based on several orthographic similarity measures that we use as features for classification. We test each feature separately, we average the values of all features, and we also explore various ways to combine the measures, by applying several Machine Learning techniques from the Weka package[1].

The task of disambiguating partial cognates can be seen as a coarse grain cross-language Word-Sense Discrimination. The results of this process can be useful for different NLP tasks and applications. Our proposed methods can be applied to any pair of languages for which a parallel corpus is available, and two monolingual collections of text. One of our main focus is to be able to disambiguate a French partial cognate looking at its English cognate and false friend senses.

[1] http ://www.cs.waikato.ac.nz/ml/weka/

We implemented a tool, a Computer-Assisted Language Learning (CALL) tool that is capable to annotate cognates and false friends in French texts, in order to help second language learners of French (native English speakers) in a reading comprehension and vocabulary retention task.

In the following sections we present related work, followed by our methods for cognate and false friend identification and its evaluation ; then we briefly describe our partial cognate disambiguation method and its evaluation, and at the end we present the CALL tool.

2 Related work

Previous work on automatic cognate identification is mostly related to bilingual corpora and translation lexicons. (Simard *et al.*, 1992) use cognates to align sentences in bitexts. They employ a very simple test : French-English word pairs are assumed to be cognates if their first four characters are identical. (Brew & McKelvie, 1996) extract French-English cognates and false friends from aligned bitexts using a variety of orthographic similarity measures based on DICE's coefficient measure. They look only at pairs of verbs in French and English, pairs that were automatically extracted from the aligned corpus.

One of the most active researchers in automatic identification of cognates between various pairs of languages is (Kondrak, 2001), (Kondrak, 2004). His work is related to the phonetic aspect of cognate identification, especially **genetic cognates** – pairs in related languages that derive directly from the same word in the ancestor (proto)-language. He uses algorithms that combine different orthographic and phonetic measures, recurrent sound correspondences, and semantic similarity based on gloss overlap.

For French and English, substantial work on cognate detection was done manually. (LeBlanc & Séguin, 1996) collected 23,160 French-English cognate pairs from two general-purpose dictionaries (Robert-Collins and Larousse-Saturne). They concluded that cognates appear to make up over 30% of the vocabulary.

Claims that false friends can be a hindrance in second language learning are supported by the studies of (Carroll, 1992). She suggests that a cognate pairing process between two words that look alike happens faster in the learner's mind than a false-friend pairing. Experiments with second language learners of different stages conducted by (Heuven *et al.*, 1998) suggest that missing false-friend recognition can be corrected when cross-language activation is used.

Word Sense Disambiguation (WSD) is an NLP task that has attracted researchers since 1950 and it is still a topic of high interest. We define the partial cognate disambiguation task as a cross-language WSD task. Determining the sense of an ambiguous word, using bootstrapping and texts from a different language was done by (Yarowsky, 1995), (Hearst, 1991), (Diab & Resnik, 2002), and (Li & Li, 2004). We follow a similar approach for our partial cognate disambiguation task. The difference between our approach and the ones mentioned above, is that our technique uses the whole sentence from the parallel text, not only the target words (the translation of certain English words), unlike (Diab & Resnik, 2002) ; we do not impose any constrains like (Yarowsky, 1995), our focus is not only on nouns as in (Hearst, 1991) ; and we look at words that are difficult to disambiguate even for humans, not only at very different words as in (Li & Li, 2004).

3 Cognates and false friends identification

3.1 Data sets for cognate and false friend identification

The data sets that we used to perform experiments for the task of cognates and false friends identification consist in a training and a testing list of pairs of words that are manually annotated as being cognates or false friends. The training dataset that we used contains 1454 pairs of French and English words.

They were extracted from different resources[2]. A separate test set composed of 1040 pairs were also extracted. A summary of the data that we have used is presented in Table 1.

	Training set	Test set
Cognates	613 (73)	603 (178)
False Friends	314 (135)	94 (46)
Unrelated	527 (0)	343 (0)
Total	1454	1040

TAB. 1 – The composition of data sets. The numbers in brackets are counts of word pairs that are identical, ignoring accents.

3.2 Method for cognate and false friend identification

Our contribution to the task of identifying cognates and false friends between languages is the method itself, the way we approach the identification task by using ML techniques. Other methods that have been proposed for cognate and false friend identification require intensive human knowledge (Barker & Sutcliffe, 2000). We used different supervised ML algorithms to best discriminate between the two classes that we have chosen : Cognates/False Friends — are orthographically similar, and Unrelated — are not orthographically similar.

In our method an instance is a pair of words containing a French word and an English word. The features that we have chosen to use in our method are the 13 orthographic similarity measures. We performed experiments when we use different feature combinations, each pair of words is represented by all 13 orthographic similarity measures, by one of the measures (we want to determine a threshold value for each measure) or by the average result of all measures. No matter what are the features that the method uses, the values of the features are real numbers between 0 and 1 (inclusively) that reflect the orthographic similarity between two words from a French-English pair (see (Inkpen *et al.*, 2005) for a detailed description of the measures).

3.3 Evaluation results for cognate and false friend identification

We present evaluation experiments using the two datasets described in Section 3.1. We classify a pair of words on the basis of similarity into two classes : Cognates/False-Friends and Unrelated. Cognates are later distinguished from false friends by virtue of being mutual translations. We report the accuracy values for the classification task (the precision and recall values for the two classes are similar to the accuracy values).

[2]See (Inkpen *et al.*, 2005) for a description of these resources.

Results on the test data set. Table 2 presents the results that we obtained for the cognate and false friend identification task on the test set. We report results for each measure separately, the average of all measures and when we used all 13 measures as features for ML algorithms.

For each measure, we need to choose a specific similarity threshold for separating Cognates/False-Friends from the Unrelated pairs. The separation has to be made such that all the pairs with similarity above or equal to the threshold are classified as Cognates/False-Friends, and all the pairs with similarity below the threshold are classified as Unrelated. We determined the best thresholds by running Decision Stump classifiers with a single feature. Decision Stumps are Decision Trees that have a single node containing the feature value that produces the best split.

We also trained several machine learning classifiers from the Weka package : OneRule (a shallow Decision Rule that considers only the best feature and several values for it), Naïve Bayes, Decision Trees, Instance-based Learning (IBK), Ada Boost, Multi-layered Perceptron, and a light version of Support Vector Machine (SMO) to experiment our method with all 13 measures. Surprisingly, only the Naïve Bayes classifier outperforms the simple average of orthographic measures. Among the individual orthographic measures, XXDICE performs the best, supporting the results on French-English cognates reported in (Brew & McKelvie, 1996).

We run similar experiments on the training data using 10-fold cross validation and we obtained similar results (better with 1-2%) as the ones that we obtained on the test set. Overall, the measures that performed best on the training set achieve more than 93% on the test set. We conclude that our classifiers are generic enough : they perform very well on the test set.

Results for three-class classification. We also experiment our method by adding one more feature and increasing the number of classes to three. The new feature which is set to 1 if the two words are translations of each other, and to 0 otherwise. The three classes are : Cognates, False-Friends and Unrelated.

As expected, this experiment achieved similar but slightly lower results than the ones from Table 2. Most of the machine learning algorithms (except the Decision Tree) did not perfectly separate the Cognate/False-Friends class. We conclude that it is better to do the two-way classification that we presented above (into Cognates/False-Friends and Unrelated), and then split the first class into Cognates and False-Friends on the basis on the value of the translation feature taken from a bilingual dictionary or a bilingual list of words.

Error analysis. We examined the misclassified pairs for the classifiers built on the training data. There were many shared pairs among the 60–70 pairs misclassified by several of the best classifiers. Several of the measures are particularly sensitive to the initial letter of the word, which is a strong clue of cognation, *arrêt - arm*, *peine - pear*. Also, the presence of an identical prefix made some pairs look similar, but they are not cognates unless the word roots are related.

3.4 Complete lists of cognates and false friends between two languages

We applied the methods of classifying pairs of words into cognates or false friends to the task of creating complete lists of cognates and false friends between two languages. As knowledge of which pairs are translations we use a bilingual dictionary and a bilingual list of words.

Classifier (measure or combination) set	Accuracy on test set
IDENT	55.00%
PREFIX	90.97%
DICE	93.37%
LCSR	94.24%
NED	93.57%
SOUNDEX	84.54%
TRI	92.13%
XDICE	94.52%
XXDICE	95.39%
BI-SIM	93.95%
BI-DIST	94.04%
TRI-SIM	93.28%
TRI-DIST	93.85%
Average measure	94.14%
Baseline	66.98%
OneRule	92.89%
Naive Bayes	94.62%
Decision Trees	92.08%
DecTree (pruned)	93.18%
IBK	92.80%
Ada Boost	93.47%
Perceptron	91.55%
SVM (SMO)	93.76%

TAB. 2 – Results on the test set of the classifiers built on the training set (individual measures and machine learning combinations).

Method description. For each pair of words that have high ortographic similarity, if they are transation of each other we put the pair in the list of cognates, otherwise we put it in the list of false friends.

For these experiments we used the XXDICE measure with a threshold of 0.14. The reason why we have chosen to use this measure is that it was the one that performed best on the test set (see Table 2). The threshold automatically determined by the method in the previous section was 0.12. We increased it a little because we wanted to obtain pairs that are classified with a higher confidence.

Results for building large lists of cognates and false friends from dictionary entry lists. To collect pairs of words that are translations of each other we used the dictionary entries from the *Internet Dictionary Project* (IDP)[3]. From the 3,246 dictionary entries we extracted 2,591 entries that were not multi-word expressions. The dictionary is not very big but it is one of the few that has its entries available for free download. We wanted to perform experiments with dictionary entries to see what percentage of the entries is recognized as cognates by our method. We concluded that 55% of the dictionary entries are classified as cognates.

[3]http ://www.june29.com/IDP/IDPfiles.html

To determine pairs of words that are not translations of each other — possible false friends, we paired each entry word with all others except its translation. Using this approach we obtained a list of 5,619,270 pairs of words that are not translations of each other. From the total number of pair of words that we created and are not translation of each other only 2% were determined to be orthographically similar enough to be false friends.

Results for building large lists of cognates and false friends from monolingual lists of words. In order to produce complete lists of words between two languages, we used the English entries from the LDOCE[4] dictionary, which is a dictionary intended for adult learners of English. We extracted 38,768 entries, and paired each entry with a list of 65,000 lemmas of French content words (nouns, adjectives, verbs, and adverbs) from the *Analyse et Traitement Informatique de la Langue Française* (ATILF[5]) project. After we paired each English word with each French word we obtained a list of pairs of words that we tried to classify in cognates and false friends using an on-line French-English Dictionary[6] of approximately 75,000 terms. From all pairs of words that were created we selected only the ones that have an XXDICE orthographic similarity value greater than 0.14. The number of pairs that are selected as similar is 11,469,662. From this number, only 3,496 pairs were identified as cognates and 3,767,435 as false friends.

As mentioned, one of our goals is to be able to produce complete list of cognates and false friends to be used in CALL tools. The pairs that we determine are not 100% accurate — they are produced automatically, they could be if validated by a human judge. This would require significantly less effort than manually building the lists from scratch. If we look at the way we determine the cognate and false friends we see that we are close to 100% recall ; we might miss the genetic cognates that have a common origin and but changed their spelling significantly.

4 Partial cognate disambiguation

This section presents our proposed techniques, based on Machine Learning, to disambiguate partial cognates. Partial cognates behave as cognates in some contexts, and as false friends in others. We use a semi-supervised method based on Monolingual and Bilingual Bootstrapping and parallel corpora to automatically create and tag our training seeds for the bootstrapping techniques. In addition to all the methods that use bootstrapping and parallel text, we also bootstrap our method with corpora from different domains. Our method uses a small set of seeds from Hansard, but additional knowledge from different domains is added using bootstrapping.

We use a supervised method to train classifiers on a part of automatically annotated data (collected from Hansard and EuroParl) and test their performance on the part of the data set aside for testing. We use 2/3 of the automatically tagged data as training and the 1/3 part for testing, an average of 130 sentences for the cognate class for training and 66 for testing an average of 100 sentences for the false friend class for training and 50 for testing. We used a set of 10 French partial cognates for which we had the corresponding English cognate and false friend words.

For the supervised method we obtained an average of 80% accuracy in disambiguating sentences that contain a French partial cognate. The best classifier among several was Naïve Bayes.

[4]http ://www.longman.com/ldoce/
[5]http ://actarus.atilf.fr/morphalou/
[6]http ://humanities.uchicago.edu/orgs/ARTFL/forms_unrest/FRENG.html

Its results are much higher than the baseline of 59% when choosing the most frequent class.

For the semi-supervised methods, we use unlabeled data from the LeMonde and the Hansard corpus and we obtained an increase in accuracy of 2%. Statistically significant improvements were obtained when we performed experiments combining corpora from different domains. More detailed results for all these experiments can be found in (Frunza & Inkpen, 2006).

5 A tool for cross-language pair annotations

In this section, we describe our tool called Cross-Language Pair Annotator (CLPA) that is capable of automatically annotating cognates and false friends in French texts. The tool uses the Unstructured Information Management Architecture (UIMA)[7] Software Development Kit (SDK) from IBM and the Baseline Information Extraction (BaLIE)[8], an open source Java project capable of extracting information from raw texts. CLPA is a tool that has a Graphical User Interface (GUI) capability that makes it easy for the user to distinguish between different annotations of the text. We designed the tool as a Java open source downloadable kit that contains all the additional projects (Balie and UIMA). Our tool is a practical follow up to the research that we did on cognates and false friends between French and English and it has the goal to help second-language learners of French. In its first version, the CLPA tool uses as knowledge a list of 1,766 cognates and a list of 428 false friends. The list of false friends contains a French definition for the French word and an English definition for the English word of the pair. Both lists contain the cognates and false friend pairs that were used in the Machine Learning experiments for the cognate and false friend identification task described in Section 3.

UIMA is an open platform for creating, integrating, and deploying unstructured information management solutions from a combination of semantic analysis and search components. It offers CLPA the GUI interface and an efficient management of the annotations that are done for a certain text. The user can select/deselect the cognate or false friend annotations. By default, both type of cross language pairs are annotated. BaLIE is a trainable Java open source project that can perform : Language Identification, Sentence Boundary Detection, Tokenization, Part of Speech Tagging and Name Entity Recognition for English, French, German, Spanish and Romanian. We use BaLIE for the tokenization and part-of-speech tagging capabilities.

The annotations that the tool makes are only for French content words : nouns, adjectives, adverbs and verbs. We have chosen to annotate only the content words to not introduce some false alarms (e.g. the French word *pour* can be either adverb (*pro*), or preposition (*for ; to*), and it is a false friend with the English word *pour* that is a verb), and also because they are of more interest for second language learners.

The user can click on one of the text annotations in the GUI to obtain additional information about the chosen annotation, (e.g. at what position in the text does the chosen word starts, what position does it end, the French definition of the French false friend word, the English definition of the English false friend word, etc.) Snapshots of the tool along with the tool itself free to download can be found here[9].

[7] http ://www.research.ibm.com/UIMA/

[8] http ://balie.sourceforge.net/

[9] www.site.uottawa.ca/∼ofrunza/Pages/CLPA.html

6 Conclusions and future work

In Section 3, we presented and evaluated a new method of identifying cognates and false friends between French and English. The method uses 13 orthographic similarity measures that are combined through different ML techniques. For each measure we determined a threshold of orthographic similarity that can be used to identify new pairs of cognates and false friends. The novelty that we bring to this task is the way we use and combine different orthographic similarity measures and the results show that the method can be used with success.

In addition to the ML technique that identifies cognates and false friends, we proposed a method that uses a bilingual dictionary to create complete lists of cognates and false friends between two languages. For highly accurate results, the human effort that is needed is significantly lower than in the case of using only human knowledge, as done in previous work.

For the task of partial cognate disambiguation (Section 4) we proposed a method that has a pure ML supervised approach and a semi-supervised method that contains two algorithms : Monolingual and Bilingual Bootstrapping, which use free unlabeled texts. Our results show that simple methods and freely available tools lead to good results and cope well with the noise that might be present in the data, in a task that is hard to solve even for humans.

In the Section 5 we presented a CALL tool, CLPA, which is able to annotate cognates and false friends in a French text. CLPA has an easy to use GUI that allows users to choose between annotations —only cognate annotation, only false friend annotation, or both, and also provide additional information to the users. This information can be useful to a second language learner similar to the feedback from a tutor.

In future work we want to apply the cognate and false friend identification task to other pairs of languages that lack this kind of resource (since the orthographic similarity measures are not language-dependent). We want to increase the accuracy of the automatically generated lists of cognates and false friends by increasing the threshold used — we could obtain better precision but less recall for both classes. We could eliminate some falsely determined false friends by using other orthographic measures or the same measure with a higher threshold on the initial list — determined with the same threshold for both classes. For the disambiguation task, we want to look at different data representations, use lemmatization and POS tagging, and apply our method to new pairs of languages (all we need is a parallel corpus, and monolingual corpora). We also want to continue to develop the tool, add other features, perform the lemmatization step, and also annotate partial cognates with the corresponding meaning in the texts.

The overall contribution of this paper is the new methods that we proposed, experimented and evaluated, and the new directions that we followed for cognate, false friend, and partial cognate words between French and English.

References

BARKER G. & SUTCLIFFE R. F. E. (2000). *An Experiment in the Semi-Automatic Identification of False-Cognates between English and Polish*. Rapport interne, Department of Languages and Cultural Studies, University of Limerick, Ireland.

BREW C. & MCKELVIE D. (1996). Word-pair extraction for lexicography. In *Proceedings of 2nd International Conf. on New Methods in Language Processing*, p. 45–55, Ankara, Turkey.

CARROLL S. (1992). *On Cognates*. Rapport interne, Second Language Research.

DIAB M. & RESNIK P. (2002). An unsupervised method for word sense tagging using parallel corpora. *In Proceedings of Association for Computational Linguistics (ACL '02)*, p. 255–262.

FRUNZA O. & INKPEN D. (2006). Semi-supervised learning of partial cognates using bilingual bootstrapping. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, COLING-ACL 2006*, p. 433–440, Sydney, Australia.

GASS S. (1987). The use and acquisition of the second language lexicon. *Studies in Second Language Acquisition 9(2)*, **9**, 128–262.

HEARST M. (1991). Noun homograph disambiguation using local context in large corpora. *Proceedings of the 7th Annual Conf. of the University of Waterloo Centre for the New OED and Text Research*, p. 1–19.

HEUVEN W. V., DIJKSTRA A. & GRAINGER J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, **39**, 458–483.

INKPEN D., FRUNZA O. & KONDRAK G. (2005). Automatic identification of Cognates and False Friends in French and English. In *RANLP-2005*, p. 251–257, Bulgaria.

KONDRAK G. (2001). Identifying Cognates by Phonetic and Semantic Similarity. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, p. 103–110.

KONDRAK G. (2004). Combining evidence in cognate identification. In *Proceedings of Canadian AI 2004 : 17th Conference of the Canadian Society for Computational Studies of Intelligence*, p. 44–59.

LEBLANC R., COMPAIN J., DUQUETTE L. & SÉGUIN H. (1989). *L'enseignement des langues secondes aux adultes : recherches et pratiques*. Les Presses de l'Université d'Ottawa.

LEBLANC R. & SÉGUIN H. (1996). Les congénères homographes et parographes anglais-français. In *Twenty-Five Years of Second Language Teaching at the Univ. of Ottawa*, p. 69–91.

LI H. & LI C. (2004). Word translation disambiguation using bilingual bootstrap. *Computational Linguistics*, **30**(1), 1–22.

MARCU D., KONDRAK G. & KNIGHT K. (2003). Cognates can improve statistical translation models. *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, p. 46–48.

MELAMED I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, **25**, 107–130.

SIMARD M., FOSTER G. F. & ISABELLE P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, p. 67–81, Montreal, Canada.

YAROWSKY D. (1995). Unsupervised word sence disambiguation rivaling supervised methods. *In Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL-95)*, p. 189–196.

Enrichissement d'un lexique bilingue par analogie

Philippe LANGLAIS, Alexandre PATRY
Université de Montréal
CP. 6128 succursale centre-ville
{felipe,patryale}@iro.umontreal.ca

Résumé. La présence de mots inconnus dans les applications langagières représente un défi de taille bien connu auquel n'échappe pas la traduction automatique. Les systèmes professionnels de traduction offrent à cet effet à leurs utilisateurs la possibilité d'enrichir un lexique de base avec de nouvelles entrées. Récemment, Stroppa et Yvon (2005) démontraient l'intérêt du raisonnement par analogie pour l'analyse morphologique d'une langue. Dans cette étude, nous montrons que le raisonnement par analogie offre également une réponse adaptée au problème de la traduction d'entrées lexicales inconnues.

Abstract. Unknown words are a well-known hindrance to natural language applications. In particular, they drastically impact machine translation quality. An easy way out commercial translation systems usually offer their users is the possibility to add unknown words and their translations into a dedicated lexicon. Recently, Stroppa et Yvon (2005) shown how analogical learning alone deals nicely with morphology in different languages. In this study we show that analogical learning offers as well an elegant and efficient solution to the problem of identifying potential translations of unknown words.

Mots-clés : analogie formelle, enrichissement de lexiques bilingues, traduction automatique.

Keywords: formal analogy, bilingual lexicon projection, machine translation.

1 Introduction

Le raisonnement par analogie est un principe bien connu en sciences cognitives et en intelligence artificielle (Gentner *et al.*, 2001). L'aptitude à raisonner par analogie a longtemps fait l'objet de questions dans les tests SAT (Scholastic Assessment Test) aux États-Unis[1]. Turney et Littman (2005) décrivent une approche basée sur le modèle de l'espace vectoriel populaire en recherche d'information qui permet de répondre à 47% de 374 questions typiquement posées dans ces tests.

On trouve dans (Lepage, 2003) un traitement particulièrement riche du rôle de l'analogie dans la langue, tant d'un point de vue formel, algorithmique qu'historique. L'auteur décrit différentes expériences, notamment en analyse morphologique, qui attestent du bien-fondé applicatif de

[1] Les tests SAT introduits en 1926 aux États-unis incluaient des tests analogiques qui ont été retirés en 2005 (http://www.collegeboard.com/about/newstat/newstat.html).

l'analogie. Des principes dégagés dans ce travail, Lepage et Denoual (2005) présentaient récemment le système ALEPH, un système de traduction par l'exemple entièrement basé sur le principe de résolution d'analogies formelles. Une *analogie formelle* met en relation quatre entités, ce que l'on dénote $[\texttt{A} : \texttt{B} = \texttt{C} : \texttt{D}]$ et qui se lit: "A est à B ce que C est à D". Ce système d'une élégance remarquable[2] montrait des performances état-de-l'art dans les tâches partagées organisées en marge des ateliers IWSLT[3].

Stroppa et Yvon (2005) proposent une formalisation algébrique à la fois concise et accessible de l'analogie formelle et décrivent les fondements théoriques de l'apprentissage analogique. Ils démontrent expérimentalement l'élégance et la puissance de cette approche dans deux tâches d'étiquetage morphologique ; la première consistait à étiqueter morpho-syntaxiquement à l'aide d'un jeu d'étiquettes fines les mots inconnus d'une langue ; la seconde visait à prédire l'arbre d'analyse morphologique d'un lemme inconnu (lire également (Yvon *et al.*, 2004)).

D'autres auteurs se sont intéressés ces dernières années au potentiel applicatif des analogies formelles. Claveau et L'Homme (2005) montraient notamment qu'un type particulier d'analogies formelles très simples à calculer permettait de structurer les termes d'un domaine. Moreau et Claveau (2006) montrent également le bénéfice du raisonnement par analogie pour l'extension de requêtes dans un système monolingue de recherche d'information.

Dans cette étude, nous montrons que le raisonnement par analogie offre également une réponse adéquate au problème concret de la traduction d'entrées lexicales inconnues. Nous rappelons en section 2 le principe général de l'apprentissage analogique. Nous présentons en section 3 comment il peut être appliqué à la tâche d'enrichissement d'un lexique bilingue. Nous évaluons notre approche en section 4 en la comparant à deux systèmes de base. Nous montrons que notre approche permet de traduire automatiquement 60% des mots inconnus d'une application. Nous dressons en section 5 un bilan de ce travail, et proposons des perspectives de recherche.

2 Raisonnement analogique

2.1 Apprentissage analogique

L'approche mise en place dans cette étude pour l'enrichissement de lexiques bilingues s'inscrit dans le cadre théorique de l'apprentissage par analogie proposé dans (Stroppa & Yvon, 2005). Un ensemble d'apprentissage $\mathcal{L} = \{L_1, \ldots, L_N\}$ est composé de N observations. Un ensemble de traits calculés sur une observation incomplète X définit un espace d'entrée. La tâche d'inférence consiste à prédire les traits manquants de X qui définissent à leur tour un espace de sortie. On désigne par $I(X)$ et $O(X)$ les projections respectives dans l'espace d'entrée et de sortie de l'observation X. La procédure d'inférence met en œuvre trois étapes:

1. construire $\mathcal{E}_{\mathcal{I}}(X) = \{(A, B, C) \in \mathcal{L}^3 \,|\, [\texttt{I(A)} : \texttt{I(B)} = \texttt{I(C)} : \texttt{I(X)}]\}$, l'ensemble de triplets analogiques de X, également dénommés *stems* dans la suite
2. construire $\mathcal{E}_{\mathcal{O}}(X) = \{Y \,|\, [\texttt{O(A)} : \texttt{O(B)} = \texttt{O(C)} : \texttt{Y}], \forall (A, B, C) \in \mathcal{E}_{\mathcal{I}}(X)\}$ l'ensemble des solutions trouvées aux équations analogiques formées par projection des triplets analogiques de X dans l'espace de sortie.
3. choisir $O(X)$ parmi les éléments de $\mathcal{E}_{\mathcal{O}}(X)$

[2]Ce système ne fait appel à aucune distance ni à aucun seuil pour rapprocher différents exemples.
[3]http://www.slc.atr.jp

Cette procédure d'inférence partage les avantages et les inconvénients de l'approche des k plus proches voisins (k-ppv). Il s'agit en effet d'une approche d'apprentissage passive qui n'effectue aucune généralisation à partir du corpus d'entraînement qui doit donc être conservé. Contrairement au k-ppv, l'étape 1 de recherche des exemples "proches" ne requiert pas la définition d'une distance entre deux exemples mais émerge du seul principe de *commutation linguistique* (Lepage, 2003). Cette pureté a un coût: la recherche d'exemples proches est une opération de complexité cubique en N, le nombre d'exemples dans $\mathcal{L}$, alors qu'elle est seulement linéaire en N dans le cas des k-ppv. Dans de nombreuses applications incluant celle présentée dans cette étude, cette recherche est trop coûteuse pour être effectuée au complet et des heuristiques doivent être appliquées pour réduire l'espace de recherche (voir section 3.2).

Le succès de l'approche repose presque entièrement sur le concept d'équation analogique que nous décrivons ci-après ainsi que sur l'hypothèse qu'il existe une correspondance entre les analogies construites sur l'espace d'entrée et leur projection dans l'espace de sortie.

2.2 Équation analogique

Différents niveaux paradigmatiques peuvent unir quatre objets en relation analogique. Ainsi des relations d'ordre sémantique comme celles utilisées dans les tests SAT peuvent être décrites: [hache : bûcheron = roulette : dentiste] et [aluminium : metal = novel : book]. Des analogies dites formelles avec lesquelles nous travaillons dans cette étude dénotent des relations graphiques entre les formes mises en relation. [fournit : fleurit = fournie : fleurie] et [abandoning : abandonment = amending : amendment] sont deux exemples (l'un en français, l'autre en anglais) de relations de nature morphologique. Le lecteur intéressé trouvera dans (Lepage, 2003) de nombreux exemples de relations analogiques dans des langues très différentes. L'équation analogique [A : B = C : ?] dénote l'ensemble des formes qui sont en relation analogique avec le triplet (ou stem) $\langle \mathtt{A}, \mathtt{B}, \mathtt{C} \rangle$:

$$[\mathtt{A} : \mathtt{B} = \mathtt{C} : \ ?] = \{X \mid [\mathtt{A} : \mathtt{B} = \mathtt{C} : \mathtt{X}]\}$$

Stroppa et Yvon (2005) montrent qu'il est possible de calculer les solutions d'une équation analogique formelle à l'aide d'un transducteur à états finis. Cette approche généralise l'algorithme proposé initialement par Lepage (1998) qui réside dans la synchronisation de deux tables d'éditions: l'une entre A et B, l'autre entre A et C. Intuitivement, cet algorithme compose dans le bon ordre les sous-séquences de B et de C qui ne sont pas dans A.

Dans ce travail, nous avons implémenté une variante de cet algorithme qui calcule premièrement les deux tables d'édition[4] (opération de complexité quadratique avec la longueur, comptée en caractères, des chaînes en présence), puis qui synchronise ensuite les deux tables (opération de complexité linéaire) pour tout chemin d'édition minimal de chaque table. Comme le nombre de chemins d'édition de coût minimal peut être exponentiel, nous considérons au plus les M premiers chemins de chaque table (dans nos expériences, M a été fixé expérimentalement à la valeur non critique de 20) et un total de M^2 paires de chemins est donc au plus synchronisé.

Il est important de noter qu'une équation peut admettre zéro, une ou plusieurs solutions qui ne sont pas nécessairement des formes légitimes de la langue étudiée. L'équation [fournir : fourniront = courir : ?] a par exemple pour solution couriront.

[4] Les coûts des opérations sont unitaires à l'exception de l'opération d'insertion qui est de coût nul, précisément car les caractères de B et C "insérés" sont ceux que nous désirons conserver dans les solutions.

3 Application à l'enrichissement d'un lexique bilingue

Les principes décrits dans la section précédente peuvent être appliqués au problème de l'enrichissement d'un lexique bilingue. Cette opération qui consiste à étendre un lexique existant à de nouvelles entrées présente de nombreux intérêts pratiques, notamment dans le cas de paires de langues faiblement dotées. La couverture d'un lexique, aussi grande soit-elle, n'est pas garante de son utilité. Ainsi, même pour une paire de langues largement dotée comme le français et l'anglais, n'existe-t-il pas de lexique bilingue couvrant les termes de tous les domaines de spécialité. Ceci justifie l'intérêt de travaux visant à apprendre automatiquement à traduire les termes spécifiques comme ceux du domaine médical (Claveau & Zweigenbaum, 2005).

3.1 Approche

Notre approche peut-être illustrée sur un exemple simple. Pour chercher la traduction du mot inconnu `futilité`, nous identifions des relations analogiques dans la langue source comme: [`activités : activité = futilités : futilité`]. Nous projetons (par une opération définie plus loin) ces relations en langue cible de manière à définir des équations analogiques (cibles) comme: [`actions : action = gimmicks : ?`] dont `gimmick` est une solution.

Formellement, nous disposons d'un corpus d'apprentissage $\mathcal{L} = \{\langle S_1, T_1\rangle, \ldots, \langle S_N, T_N\rangle\}$ qui réunit des paires de mots en relation de traduction. L'espace d'entrée est l'ensemble des mots de la langue source, l'espace de sortie celui des mots de la langue cible ; et on définit[5]:

$$\forall X \equiv \langle \mathtt{S}, \mathtt{T}\rangle,\ I(X) = S \text{ et } O(X) = T$$

L'enrichissement de $\mathcal{L}$ consiste pour toute forme source S inconnue de l'espace d'entrée à identifier les triplets analogiques sources qui entrent en équation analogique avec S:

$$\mathcal{E}_{\mathcal{I}}(S) = \{\langle \mathtt{i}, \mathtt{j}, \mathtt{k}\rangle, \in [1, N]^3 \mid S_i \neq S_j \neq S_k \text{ et } [S_i : S_j = S_k : S]\}$$

Chaque élément de $\mathcal{E}_{\mathcal{I}}(S)$ est ensuite projeté dans l'espace de sortie à l'aide de l'opérateur $proj_{\mathcal{L}}$ et les solutions calculées dans cet espace sont colligées dans $\mathcal{E}_{\mathcal{O}}(S)$:

$$\mathcal{E}_{\mathcal{O}}(S) = \bigcup_{\langle \mathtt{i},\mathtt{j},\mathtt{k}\rangle \in \mathcal{E}_{\mathcal{I}}(S)} \mathcal{E}_{\langle \mathtt{i},\mathtt{j},\mathtt{k}\rangle}(S)$$

où:

$$\mathcal{E}_{\langle \mathtt{i},\mathtt{j},\mathtt{k}\rangle}(S) = \{T \mid [U : V = W : T], \forall (U, V, W) \in (proj_{\mathcal{L}}(S_i) \times proj_{\mathcal{L}}(S_j) \times proj_{\mathcal{L}}(S_k))\}$$

Le mécanisme de projection que nous utilisons consiste à simplement à retourner pour une entrée source S du lexique bilingue les associations cibles (ou traductions) qui lui correspondent (une entrée S possède potentiellement plusieurs traductions dans $\mathcal{L}$):

$$proj_{\mathcal{L}}(S) = \{T \mid \langle S, T\rangle \in \mathcal{L}\}$$

[5]Par exemple, pour l'observation $X = \langle \mathtt{déjà}, \mathtt{already}\rangle$, $I(X) = \mathtt{déjà}$ et $O(X) = \mathtt{already}$.

3.2 Implémentation

Trouver l'ensemble des triplets analogiques de S est une opération trop coûteuse en temps (cubique avec la taille de l'espace d'entrée). Nous utilisons deux techniques pour réduire cette complexité. La première consiste à utiliser les équations analogiques en mode génératif: plutôt que de vérifier tous les triplets $\langle S_i, S_j, S_k \rangle$ entretenant une relation analogique avec S, nous cherchons les solutions à $[S_j : S_i = S :?]$. Il s'agit d'une méthode exacte qui repose sur la propriété (Lepage, 2003):

$$[\mathtt{A} : \mathtt{B} = \mathtt{C} : \mathtt{D}] \equiv [\mathtt{B} : \mathtt{A} = \mathtt{D} : \mathtt{C}]$$

Cette méthode, qui réduit la construction de $\mathcal{E}_\mathcal{I}(S)$ à une opération de complexité quadratique, est encore trop coûteuse. Nous appliquons donc une seconde méthode, cette fois-ci heuristique, qui consiste à ne calculer les équations analogiques que sur les seuls mots proches de S ; formellement, nous construisons $\mathcal{E}_\mathcal{I}(S)$ selon (étape 1):

$$\mathcal{E}_\mathcal{I}(S) = \{U \mid [A : B = S : U], \forall A \in v_\delta(S) \text{ et } B \in v_\beta(A)\}$$

où $v_\gamma(A)$ est une fonction de voisinage d'une lexie A de la forme:

$$v_\gamma(A) = \{B \mid f(B, A) \leq \gamma\}$$

Dans cette étude, nous avons utilisé pour fonction f la distance d'édition (Levenshtein, 1966)[6].

Nous avons mentionné qu'une équation analogique peut générer plusieurs solutions, certaines n'étant pas des formes légitimes d'une langue. Aussi, l'étape 3 du processus d'inférence consiste dans notre cas à sélectionner les solutions analogiques les plus fréquemment générées et à ne retenir que celles qui sont présentes dans un (grand) lexique monolingue cible $\mathcal{V}$. Nous avons compilé à cet effet à partir de textes variés un lexique monolingue totalisant 466 439 formes différentes. Des exemples de traductions produites par analogie sont présentés en Table 1.

4 Expériences

Nous avons réalisé nos expériences dans le cadre de la campagne d'évaluation des systèmes de traduction qui s'est tenue lors de l'atelier WMT'06 (Koehn & Monz, 2006). Dans cette tâche, les corpus d'entraînement et de test étaient constitués de textes parlementaires européens. À l'insu des équipes participantes, les organisateurs ont ajouté aux 2 000 phrases du corpus de test (corpus `domaine` dans la suite), 1064 phrases[7] hors-domaine en provenance du site internet de `Project Syndicate` (`http://www.project-syndicate.com`), une organisation sans but lucratif qui distribue des articles de revue sur des thèmes variés (politique, économie, science, etc.). Ce corpus est baptisé `hors-domaine` dans la suite.

Nos expérimentations simulent une situation typique du développement d'un système de traduction basé sur l'exemple: nous disposons d'un corpus d'entraînement bilingue sur lequel est

[6] Le lecteur attentif aura noté que nous avons plus haut (section 2.1) souligné que contrairement à l'approche des K plus proches voisins, le raisonnement par analogie ne requiert pas de distance. La distance que nous utilisons ici n'est pas constitutive de l'approche (comme c'est le cas dans les k-ppv) mais répond seulement à des considérations pratiques: nous pourrions par exemple nous en affranchir en tirant aléatoirement des triplets dans l'espace d'entrée.

[7] 30 de ces phrases contenaient des problèmes d'encodage et ont été retirées de notre étude.

source	cand	nb	(candidat, fréquence)
anti-agricole	296	5	(anti-farm,5) (anti-agricultural,3) (anti-farming,3) (anti-rural,3) (anti-farmer,3)
concentrerait	2947	7	**(concentrat,11)** (concentrate,4) (summarized,4) (summarizing,4) (concentrating,3) (focuss,3) (focus,3)
écrirait	156	4	**(writs,1)** (write,1) (writes,1) **(writ,1)**
réintégrés	2686	18	(reinstated,20) (reintegrated,17) (re-integrated,13) (re-entered,10) (reincluded,8) (reinvolved,8) (reincorporated,8) (reinserted,7) (reinstate,7) (reintegrate,6) (reinstating,4) (accomplished,3) (rebuilt,3) (reinclude,3) (rejoined,3) (reverte,2) (reintegration,2) (reintegrating,2)
galette	218	1	(pancake,13)

TAB. 1 – Exemples de traductions obtenues par projection à partir de $\mathcal{L}_{100\,000}$. cand indique le nombre de solutions analogiques cibles générées, nb indique les traductions candidates retenues une fois validées par le lexique $\mathcal{V}$. Les traductions en gras sont clairement erronées.

appris un lexique bilingue (probabiliste dans notre cas). Nous disposons de plus d'une (grande) collection de textes en langue cible que nous avons utilisée ici pour compiler le lexique monolingue cible $\mathcal{V}$ (voir la section 3.2). Afin de bien analyser les limites de l'approche, nous nous sommes concentrés sur la paire de langues français-anglais qui nous est familière[8]. Notre but est de prédire des traductions anglaises de termes français inconnus du corpus d'entraînement. Nous avons éliminé de notre étude les formes numériques (nous pouvons les traiter de manière simple).

4.1 Évaluation automatique

Évaluer la qualité de différentes variantes de notre approche nécessite le parcours de plusieurs listes de traductions. En plus de s'avérer fastidieuse, cette entreprise s'avère délicate: beaucoup de traductions produites ne sont valides que dans certains contextes seulement. Il suffit pour cela de consulter les exemples de la Table 1 pour s'apercevoir de la difficulté de la tâche.

Nous avons donc, dans un premier temps, procédé à une évaluation automatique où un lexique bilingue de référence est utilisé. Ce lexique, dénommé $\mathcal{L}_{ref}$, est obtenu par entraînement sur le corpus au complet de WMT'06 (688 000 paires de phrases) d'un modèle statistique obtenu à l'aide de la trousse à outils GIZA++ (Och & Ney, 2000)[9]. De la même manière, nous avons entraîné des modèles lexicaux $\mathcal{L}_T$ sur différentes tranches du corpus d'entraînement ($T = 5\,000$, 10 000, 100 000, 200 000 et 500 000 paires de phrases). Nous avons alors traduit à l'aide du raisonnement analogique les mots français du corpus de test de WMT'06 qui n'étaient pas présents dans les lexiques $\mathcal{L}_T$ mais présents dans le lexique de référence $\mathcal{L}_{ref}$. Une traduction candidate est considérée correcte si elle est validée par $\mathcal{L}_{ref}$.

À des fins de comparaison, deux approches de base (*baseline*) ont été testées sur la même

[8]Des résultats similaires sont observés pour la paire de langues espagnol-anglais.

[9]En pratique, pour éliminer une partie du bruit d'un lexique appris automatiquement, nous le croisons (intersection) avec un lexique résultant de l'entraînement d'un modèle lexical entraîné dans la direction opposée (anglais-français versus français-anglais).

tâche dans les mêmes conditions. La première approche (BASE1) consiste à proposer comme traduction d'un mot source inconnu, les mots cibles les plus similaires (au sens de la distance d'édition). Cette approche marchera d'autant mieux que les langues sont proches (ex. *docteur* → *doctor*). La deuxième approche (BASE2) ressemble davantage à l'approche analogique et consiste à identifier les formes sources connues du lexique $\mathcal{L}_T$ qui sont proches du mot inconnu, puis à proposer leurs traductions telles qu'indiquées par ce lexique (ex. *demanda* → *demande* → *request*). Chacune de ces approches est testée selon deux variantes. La première ($_{id}$) propose le même nombre de traductions que l'a suggéré l'approche ANALOG dans les mêmes conditions (les approches sont donc directement comparables) ; la seconde ($_{10}$) propose dix traductions pour chaque mot inconnu.

T	5 000		10 000		50 000		100 000		200 000		500 000	
	p%	r%	p%	r%	p%	r%	p%	r%	p%	r%	p%	r%
	domaine											
ANALOG	50.8	30.7	54.4	44.3	57.9	63.9	57.0	63.8	57.7	64.4	30.4	67.6
BASE1$_{id}$	31.6	30.7	32.3	44.3	24.7	63.9	20.3	63.8	20.9	64.4	8.7	67.6
BASE2$_{id}$	34.5	30.7	37.1	44.3	39.0	63.9	37.8	63.8	34.4	64.4	56.5	67.6
BASE1$_{10}$	26.7	100.0	28.3	100.0	23.9	100.0	20.0	100.0	16.6	100.0	11.8	100.0
BASE2$_{10}$	26.3	100.0	30.8	100.0	29.3	100.0	27.6	100.0	24.9	100.0	55.9	100.0
unk	[3 171 , 6.8]		[2 245 , 6.1]		[754 , 3.7]		[456 , 2.8]		[253 , 2.0]		[34 , 1.2]	
	hors-domaine											
ANALOG	52.4	28.9	54.4	42.4	51.7	68.0	53.6	73.4	55.3	79.2	43.9	86.8
BASE1$_{id}$	28.0	28.9	29.0	42.4	27.3	68.0	23.1	73.4	26.8	79.2	22.7	86.8
BASE2$_{id}$	32.9	28.9	35.0	42.4	32.5	68.0	35.9	73.4	40.8	79.2	59.1	86.8
BASE1$_{10}$	24.7	100.0	25.9	100.0	25.1	100.0	20.9	100.0	25.2	100.0	25.0	100.0
BASE2$_{10}$	21.7	100.0	26.4	100.0	27.2	100.0	29.4	100.0	33.6	100.0	57.9	100.0
unk	[2 270 , 6.2]		[1 701 , 5.5]		[621 , 3.2]		[402 , 2.4]		[226 , 1.7]		[76 , 1.3]	

TAB. 2 – Résultats de différentes méthodes d'extension de lexique en fonction de la taille du lexique à étendre (ligne du haut). Les lignes préfixées de *unk* indiquent le nombre de mots à traduire ainsi que le nombre moyen de traductions dans $\mathcal{L}_{ref}$.

Les résultats de cette évaluation automatique sont consignés en Table 2 en fonction de la taille du lexique utilisé pour la projection. Deux mesures sont rapportées: `p%` représente le pourcentage d'entrées inconnues ayant au moins une traduction valide ; `r%` indique le pourcentage de mots traduits. La première ligne de cette table indique par exemple que sur le corpus de test `domaine`, notre approche (ANALOG) propose au moins une traduction pour 30.7% des mots du corpus de test du domaine qui sont inconnus du lexique $\mathcal{L}_{5000}$. La moitié (50.8%) de ces entrées étendues contiennent une traduction valide (selon $\mathcal{L}_{ref}$).

La Table 2 appelle plusieurs commentaires. Il convient tout d'abord de garder à l'esprit que ce que nous mesurons ici est davantage l'aptitude d'une approche à reconstruire un lexique de référence à partir d'un lexique de base. Nous ne mesurons par exemple pas ici les traductions produites pour les mots inconnus du lexique de référence. Globalement, les performances de ANALOG augmentent avec la taille du lexique de base. Ceci est normal car plus ce lexique est grand, plus il contient de formes sources qui peuvent entrer en relation analogique avec un mot inconnu et plus les traductions de ces mots sont nombreuses, ce qui permet de créer davantage de relations analogiques dans la langue cible. Les mesures faites à l'aide du lexique $\mathcal{L}_{500\,000}$

sont certainement peu fiables: seulement 34 et 76 entrées sont en effet évaluées sur le jeu de test `domaine` et `hors-domaine` respectivement. On remarque également que les approches BASE1 et BASE2 sont inférieures en qualité à l'approche ANALOG (BASE1 est sans surprise la moins bonne des trois). Il est possible avec les approches de proximité d'obtenir un taux de traduction `r%` parfait, au prix d'une détérioration de la qualité des traduction produites, ce qui suggère qu'une combinaison des deux approches (comme par exemple écouter BASE2 lorsque ANALOG est silencieuse) améliorerait les performances globales.

Nous avons évalué que pour la moitié des entrées non traduites, c'est une absence de relation analogique identifiée dans l'espace cible qui aboutit à l'absence de candidat. En moyenne sur le lexique $\mathcal{L}_{100\,000}$, une entrée inconnue entre en relation analogique 988 fois du côté source, ce qui génère en moyenne 52 formes sources qui appartiennent au lexique de projection $\mathcal{L}_T$. Du côté cible, une moyenne de 99 solutions analogiques sont proposées (par forme inconnue source) ; une moyenne de 5 d'entre-elles seulement sont validées par $\mathcal{V}$ et donc considérées ici.

4.2 Évaluation manuelle

Une inspection des traductions proposées révèle certains problèmes dont cette évaluation ne rend pas compte. En particulier, certaines entrées reçoivent une traduction correcte alors que le lexique de référence est erroné ou incomplet. C'est par exemple le cas des exemples de la Figure 1. Par exemple, `circumventing` et `fellow` sont des traductions légitimes de `coutournant` et `concitoyen` respectivement. Sur les 20 premières entrées lexicales considérées erronées par notre procédure d'évaluation, 12 contenaient des traductions valides et 4 des entrées étaient mal traduites dans le lexique de référence.

contournant (*49 candidats*)
`analog` ⋄ (circumventing,55) (undermining,20) (evading,19) (circumvented,17) (overturning,16) (circumvent,15) (circumvention,15) (bypass,13) (evade,13) (skirt,12) $\mathcal{L}_{ref}$ ⋄ **skirting**, **bypassing**, by-pass, overcoming
concitoyen (*24 candidats*)
`analog` ⋄ (citizens,26) (fellow,26) (**fellow-citizens**,26) (people,26) (citizen,23) (fellow-citizen,21) (fellows,5) (peoples,3) (civils,3) (fellowship,2) $\mathcal{L}_{ref}$ ⋄ **fellow-citizens**

FIG. 1 – Les 10 meilleures traductions produites par `analog` à patir de $\mathcal{L}_{200\,000}$ pour deux mots inconnus et leurs traductions dans $\mathcal{L}_{ref}$. Les traductions en gras sont présentes dans la liste candidate et la liste de référence.

Nous avons donc entrepris une évaluation manuelle des traductions proposées par les deux approches ANALOG et BASE2 pour les 127 termes du corpus `domaine`[10] inconnus de $\mathcal{L}_{ref}$. Nous avons décidé, non sans arbitraire, d'identifier comme valide une traduction candidate dès lors qu'elle était synonyme d'une traduction possible du mot source (*citizen* était par exemple considérée comme une traduction acceptable de *concitoyen*).

75 (60%) des mots inconnus recevaient au moins une traduction valide avec la première méthode, contre 63 (50%) avec la seconde. Sur ces mots, 61 traductions (81%) étaient proposées

[10]Nous n'avons pas observé de différence notable dans l'évaluation automatique sur les mots du et hors-domaine.

en tête par ANALOG contre 22 (35%) pour BASE2. Des 52 mots n'ayant pas reçu par ANALOG de traduction satisfaisante, 38 (73%) n'ont en fait reçu aucune traduction. Ces mots sont en majorité des noms propres, des mots d'une autre langue (Latin, Grec ou Anglais) ainsi que des mots composés.

Nous concluons de cette évaluation manuelle que le raisonnement par analogie permet pour la paire de langue français-anglais de proposer une traduction valide pour 80% des entrées inconnues *simples* (c'est-à-dire excluant les noms propres, les mots d'emprunts, les mots composés et les données chiffrées) de notre jeu de test.

5 Discussion et perspectives

Nous avons montré que le raisonnement analogique permettait de proposer une traduction valide à 60% des mots inconnus d'un jeu de test particulier (soit 80% des mots simples) pour la paire de langues français-anglais. Nous menons actuellement des expériences sur d'autres paires de langues disponibles dans le cadre de la tâche partagée de WMT'06. Des taux similaires sont observés pour la paire de langue Espagnol-Anglais, alors qu'une perte d'environ 10% est observée sur la paire allemand-anglais.

En plus de prédire la traduction de mots inconnus, nous avons remarqué que cette technique peut-être utilisée pour enrichir les traductions d'entrées lexicales peu fréquentes dans le corpus d'entraînement. Ces entrées sont souvent mal traduites par les approches probabilistes. Nous avons également observé qu'il était envisageable d'appliquer `analog` à la traduction de séquences de mots. Nous planifions donc d'étudier de manière systématique l'impact de notre approche sur un système de traduction statistique basé sur les séquences de mots.

Toute analogie n'est pas bonne à faire ; aussi souhaitons-nous apprendre automatiquement à prédire la productivité d'une analogie. Ceci offrirait notamment une méthode de sélection plus efficace que la simple fréquence que nous avons considérée dans cette étude. Consulter le corpus d'entraînement lors d'une traduction n'est pas une approche satisfaisante. Nous souhaitons donc modéliser les régularités que recèlent les équations analogiques que nous formulons, dans la veine des travaux décrits par (Claveau & Zweigenbaum, 2005). Traduire un mot inconnu pourrait alors se faire par application de règles plutôt que par consultation d'exemples.

Plusieurs auteurs se sont intéressés à l'identification de traductions dans des corpus comparables, soit pour des mots simples (Fung & Yee, 1998; Rapp, 1999; Takaaki & Matsuo, 1999), soit pour des termes de spécialité (Morin & Daille, 2004). Les techniques proposées dans ces travaux peuvent être employées à l'enrichissement d'un lexique bilingue. Il convient cependant de souligner que contrairement à ces approches, les traductions que nous proposons émergent du seul principe de l'analogie. Nous ne sommes donc pas soumis au problème non trivial de l'acquisition de corpus dédiés (parallèles ou non) qui doivent contenir les mots que nous avons à traduire ainsi que leurs traductions.

Remerciements

Cette étude a largement profité de discussions que nous avons eues avec Nicolas Stroppa et François Yvon ainsi que du tutoriel donné à TALN'06 par Yves Lepage. Nous remercions les relecteurs anonymes pour la pertinence de leurs commentaires.

Références

CLAVEAU V. & L'HOMME M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie - utilisation comparée de ressources endogènes et exogènes. In *6ème rencontre de Terminologie et Intelligence Artificielle (TIA'05)*, Rouen, France.

CLAVEAU V. & ZWEIGENBAUM P. (2005). Automatic translation of biomedical terms by supervised transducer inference. In *10th Conference on Artificial Intelligence in Medicine (AIME'05)*, Aberdeen, Écosse.

FUNG P. & YEE L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of the 36th ACL*, p. 414–420, San Francisco, California.

GENTNER D., HOLYOAK K. J. & KONIKOV B. N. (2001). *The Analogical Mind*. Cambridge, MA: MIT Press.

KOEHN P. & MONZ C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, p. 102–121, New York City: Association for Computational Linguistics.

LEPAGE Y. (1998). Solving analogies on words: an algorithm. In *COLING-ACL*, p. 728–734.

LEPAGE Y. (2003). De l'analogie rendant compte de la commutation en linguistique. Mémoire d'Habilitation à diriger des recherches, Université Joseph Fourier, Grenoble I.

LEPAGE Y. & DENOUAL E. (2005). Aleph: an ebmt system based on the preservation of proportionnal analogies between sentences across languages. In *International Workshop on Statistical Language Translation (IWSLT)*, Pittsburgh, PA.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, **6**, 707–710.

MOREAU F. & CLAVEAU V. (2006). Extensions de requêtes par relations morpho-syntaxiques apprises automatiquement. In *3ème Conférence en Recherche d'Informations et Applications (CORIA'06)*, Lyon, France.

MORIN E. & DAILLE B. (2004). Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé. *TAL*, **45(3)**, 103–122.

OCH F. & NEY H. (2000). Improved statistical alignment models. In *38th annual meeting of the Association for Computational Linguistics (ACL'00)*, p. 440–447, Hongkong, China.

RAPP R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *37th annual meeting of the Association for Computational Linguistics (ACL'99)*, p. 519–526, College Park, Maryland.

STROPPA N. & YVON F. (2005). An analogical learner for morphological analysis. In *9th Conf. on Computational Natural Language Learning (CoNLL)*, p. 120–127, Ann Arbor, MI.

TAKAAKI T. & MATSUO Y. (1999). Extraction of translation equivalents from non-parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'99)*, p. 109–119, Chester, England.

TURNEY P. & LITTMAN M. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning Journal*, **60**(1-3), 251–278.

YVON F., STROPPA N., DELHAY A. & MICLET L. (2004). *Solving analogical equations on words*. Rapport interne, École Nationale Supérieure des Télécommunications, Paris, France.

Inférence de règles de réécriture pour la traduction de termes biomédicaux

Vincent CLAVEAU
IRISA - CNRS
Campus de Beaulieu
35042 Rennes cedex, France
Vincent.Claveau@irisa.fr

Résumé. Dans le domaine biomédical, le caractère multilingue de l'accès à l'information est un problème d'importance. Dans cet article nous présentons une technique originale permettant de traduire des termes simples du domaine biomédical de et vers de nombreuses langues. Cette technique entièrement automatique repose sur l'apprentissage de règles de réécriture à partir d'exemples et l'utilisation de modèles de langues. Les évaluations présentées sont menées sur différentes paires de langues (français-anglais, espagnol-portugais, tchèque-anglais, russe-anglais...). Elles montrent que cette approche est très efficace et offre des performances variables selon les langues mais très bonnes dans l'ensemble et nettement supérieures à celles disponibles dans l'état de l'art. Les taux de précision de traductions s'étagent ainsi de 57.5 % pour la paire russe-anglais jusqu'à 85 % pour la paire espagnol-portugais et la paire français-anglais.

Abstract. In the biomedical domain, offering a multilingual access to specialized information is a major issue. In this paper, we present an original approach to translate simple biomedical terms between several languages. This fully automatic approach is based on a machine learning technique inferring rewriting rules and on language models. The experiments that are presented are done onn several language pairs (French-English, Spanish-Portuguese, Czech-English, Russian-English...). They demonstrate the efficiency of our approach by yielding translation performances that vary according to the languages but are always very good and better than those of state-of-art techniques. Indeed, the translation precision rates go from 57.5 % for translation from Russian to English up to 85 % for Spanish-Portuguese and French-English language pairs.

Mots-clés : traduction artificielle, terminologie biomédicale, apprentissage artificiel, modèles de langue.

Keywords: machine translation, biomedical terminology, machine learning, language models.

1 Introduction

Dans le domaine biomédical, les problématiques d'accès à l'information sont particulièrement importants. De nombreux documents sont en effet quotidiennement collectés dans des bases

spécialisées très consultées. La base PubMed regroupe par exemple 16 millions de publications médicales et fait face à plus de 3 millions de requêtes par jour. Dans la plupart de ces bases, les documents sont indexés à l'aide de terminologies de référence en anglais ; la mise en place de stratégies multilingues pour faciliter l'accès à ces bases pour les non-anglophones est donc cruciale. Quelques terminologies biomédicales multilingues existent, mais elles sont mises en défaut par l'évolution rapide des connaissances et le manque de moyens pour certaines langues.

En réponse à ces besoins, nous présentons et évaluons dans cet article une méthode de traduction automatique de termes biomédicaux fonctionnant sur différentes langues (anglais, espagnol, français, russe...). Cette méthode permet de produire des traductions de termes simples (*i.e.* composés d'un seul mot) du domaine biomédical d'une langue source dans une langue cible. Ce travail repose sur deux hypothèses majeures : 1- dans le domaine biomédical, les termes équivalents entre deux langues sont souvent morphologiquement proches ; 2- les différences entre termes de chaque langue sont régulières et peuvent être apprises automatiquement. Ces deux hypothèses s'appuient sur le fait que les termes biomédicaux sont construits sur les mêmes racines grecques et latines (Namer, 2005), et leurs dérivations très régulières (*e.g.* pour le couple français-anglais *ophtalmorragie*/*ophthalmorrhagia*, *ophtalmoplastie*/*ophthalmoplasty*, *leucorragie*/*leukorrhagia*).

Notre approche s'appuie sur une technique d'apprentissage artificiel simple que nous avons développée. Elle nous permet d'inférer un ensemble de règles de réécriture à partir de couples de termes langue source-langue cible traductions l'un de l'autre. Ces règles, une fois apprises, peuvent alors être appliquées à des termes de la langue source pour produire les termes équivalents dans la langue cible. Il est intéressant de noter qu'aucune connaissance, ni intervention humaine n'est requise à l'exception de la phase de supervision (*i.e.* la constitution de l'ensemble de couples de termes langue source-langue cible), celle-ci pouvant se faire simplement en exploitant les terminologies multilingues existantes.

Dans la section suivante, nous présentons les travaux connexes à notre problématique. Nous décrivons ensuite en section 3 notre technique de traduction de termes biomédicaux que nous évaluons sur différentes paires de langues en section 4.

2 Travaux connexes

Peu de travaux se placent dans le cadre de la traduction directe de termes, et moins encore dans le domaine biomédical. Cette problématique a cependant déjà été abordée et une solution fonctionnelle a été proposée par Claveau & Zweigenbaum (2005b; 2005a). Celle-ci repose sur une technique d'apprentissage de transducteurs mais ne peut s'appliquer qu'aux langues partageant le même alphabet, contrairement à l'approche présentée ici. Outre sa plus grande souplesse, nous montrons en section 4 que notre technique obtient, de plus, des performances supérieures à cette approche par transducteurs. Schulz *et al.* (2004) ont également proposé une technique de traduction de termes biomédicaux, du portugais vers l'espagnol, s'appuyant sur des règles de réécriture. Cependant, ces règles sont fournies manuellement ; une telle approche n'est donc pas envisageable à grande échelle pour le traitement de plusieurs paires de langues.

Hors du domaine biomédical, des problématiques proches sont parfois abordées dans le domaine de la traduction automatique de textes. Ainsi, la détection de cognats (couples de mots bilingues de formes proches) (Fluhr *et al.*, 2000, *inter alia*) s'appuie sur des opérations morphologiques simples parfois proches des règles de réécriture que nous inférons. D'autres travaux

reposent quant à eux sur des recherches en corpus à l'aide de techniques statistiques de cooccurrences pour trouver des alignements – et donc des relations de traduction potentielle – entre termes dans des corpus alignés (Ahrenberg *et al.*, 2000; Gale & Church, 1991) ou comparables (Fung & McKeown, 1997). Outre le problème de la rareté de corpus spécialisés alignés, ces approches diffèrent de la nôtre en cela qu'il s'agit pour ces auteurs de retrouver une traduction d'un mot dans un texte (mise en relation), alors que nous nous posons dans le cadre plus strict de la traduction (génération). Mentionnons enfin les travaux sur la translittération, notamment du katakana ou de l'arabe (Tsuji *et al.*, 2002; Knight & Graehl, 1998, par exemple). Les techniques utilisées dans ceux-ci sont parfois proches de celle proposée ici, mais ne concernent que la représentation d'imports dans des langues ayant un alphabet différent de la langue source.

3 Technique de traduction artificielle de termes

La technique de traduction de termes biomédicaux que nous proposons fonctionne en deux temps. Tout d'abord, des règles de réécriture sont inférées à partir d'exemples de paires de termes traductions l'un de l'autre (sous-section 3.1). Un modèle de langue est ensuite appris et utilisé pour choisir la traduction la plus probable parmi les possibilités générées par l'application des règles de réécriture inférées (sous-section 3.2). Nous terminons la section par quelques commentaires sur cette technique de traduction de termes.

3.1 Inférence de règles de réécriture

La technique de traduction proposée repose sur l'apprentissage de règles de réécriture (que l'on peut aussi voir comme des règles de translittération). Ces règles, apprises à partir de listes de paires bilingues de termes du domaine (*cf.* section 4.1), sont de la forme : $\langle input\ string \rangle \rightarrow \langle output\ string \rangle$. Dans la suite de l'article, nous notons r une règle de réécriture, $\mathcal{R}$ est la liste de toutes les règles inférées pendant une expérience, $input(r)$ et $output(r)$ désignent respectivement la chaîne d'entrée et la chaîne de sortie de la règle r.

Algorithme 1 Apprentissage des règles de réécriture

```
1: aligner les paires de termes au niveau des lettres, mettre le résultat dans L
2: for all paire de termes W1 dans L do
3:     for all alignement de lettres dont les 2 lettres diffèrent dans W1 do
4:         trouver la meilleure hypothèse de règles r dans l'espace de recherche E
5:         ajouter r à l'ensemble de règles R
6:     end for
7: end for
```

L'algorithme 1 donne un aperçu global de notre technique d'apprentissage. La première étape est réalisée à l'aide de DPalign (`http://www.cnts.ua.ac.be/~decadt/?section=dpalign`). Ce logiciel aligne deux séquences en minimisant leur distance d'édition par programmation dynamique selon l'algorithme de Wagner & Fischer (1974) ; les coûts de substitution des caractères sont calculés sur l'ensemble des paires à aligner. Ce logiciel ne repose donc pas sur une similarité formelle des caractères pour aligner les séquences ; il nous est ainsi possible d'aligner des termes ne partageant pas le même alphabet.

Une liste de paires de termes est donnée en entrée ; à chaque terme sont ajoutés deux caractères # pour représenter le début et la fin de la chaîne de caractères. La liste de sortie $\mathcal{L}$ va alors contenir les paires de termes alignés au niveau des lettres ; le tableau 1 en présente quelques exemples ('_' signifie *aucun caractère*). Par la suite, le terme d'entrée (respectivement de sortie) d'une telle paire alignée p est noté $input(p)$ (resp. $output(p)$) ; de plus, $align(x, y)$ indique que la sous-chaîne x est alignée avec la sous-chaîne y dans la paire de termes considérée.

$\mathcal{L}$ portugais-anglais	$\mathcal{L}$ anglais-russe
#cetosteróides#	#adenosinetriphosphatase#
#ketosteroid_s#	#аденозин_трифосфатаза#
#electroporação_#	#hydroxypregnenolone#
#electroporation#	#гидроксипрегненолон_#
#encef_alograf_ia#	#keratoplasty__#
#encephalography_#	#кератопластика#

TAB. 1 – Exemples d'alignements produits pour deux paires de langues

Dans notre processus d'apprentissage, ces paires de mots alignés sont considérées comme des exemples à partir desquels les deux boucles imbriquées infèrent des règles de réécriture. Comme pour beaucoup de problèmes d'apprentissage artificiel symbolique, cette phase d'inférence (ligne 4) peut être considérée comme un problème de parcours d'espace. À chaque élément de cet espace est associé un score ; on cherche à trouver l'élément de l'espace maximisant ce score. Dans notre cas, l'espace de recherche est composé de toutes les règles de réécriture possibles compatibles avec l'exemple choisi. Par exemple, considérons que la paire de mots W1 choisie à la ligne 2 est *#opht_almologie#/#ophthalmology_#*, et supposons que c'est l'alignement *i/y* qui est choisi à la ligne 3. Quelques règles de réécriture compatibles dans ce contexte sont *i* $\rightarrow$ *y*, *gi* $\rightarrow$ *gy*, *ie* $\rightarrow$ *y* (on ne note pas le caractère _ dans les règles), *ologie#* $\rightarrow$ *ology#*...

Le score d'une règle est calculé à partir de la liste $\mathcal{L}$; c'est le ratio entre le nombre de fois où la règle s'applique aux termes alignés de la liste d'exemples et le nombre de fois où la prémisse de la règle apparaît dans les termes source de la liste d'exemples. Formellement, le score d'une règle r est donc défini par ($\subseteq$ représente l'inclusion de chaîne de caractères) :

$$score(r) = \frac{|\{p \in \mathcal{L} \mid input(r) \subseteq input(p) \wedge output(r) \subseteq align(input(r), p)\}|}{|\{s \in \mathcal{L}_{input} \mid input(r) \subseteq s\}|}$$

Du fait du très grand nombre de règles possibles, chercher la règle maximisant la fonction de score pour chacun des exemples peut être une tâche très lourde en temps de calcul. Heureusement, l'espace de recherche peut être organisé hiérarchiquement pour rendre l'exploration plus efficace. En effet, les règles compatibles pour un exemple peuvent être organisées de la plus générale à la plus spécifique avec la notion de subsomption suivante :

$$r_1 \succeq r_2 \Leftrightarrow (input(r_1) \subseteq input(r_2) \wedge output(r_1) \subseteq output(r_2)).$$

Cette relation de subsomption est réflexive, antisymétrique et transitive ; l'espace résultant est un treillis. La figure 1 présente l'espace de recherche organisé par cette subsomption construit à partir de l'exemple *i/y* dans *#opht_almologie#/#ophthalmology_#*. Dans notre cas, la recherche est effectuée du plus général au plus spécifique (*top-down*) ; cela, et les propriétés d'héritage que cette structure implique, nous permet de rechercher efficacement la meilleure règle (calcul du score d'une règle en n'examinant que les paires de termes que son père couvre, élagage de l'espace basé sur le meilleur score courant...).

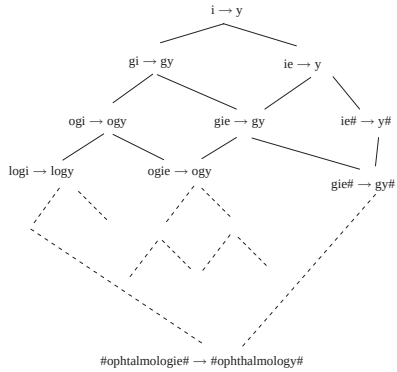


FIG. 1 – Treillis de recherche de l'exemple *i/y* dans *#opht_almologie#/#ophthalmology_#*

3.2 Évaluation des traductions proposées

L'algorithme présenté ci-dessus va potentiellement générer une règle de réécriture par différence pour chacune des paires de termes utilisées en exemple, conduisant à obtenir un grand nombre de règles. Pour traduire un terme inconnu dans la langue d'entrée, toutes les règles applicables à ce terme (*i.e.* les règles dont la prémisse correspond à une sous-chaîne du terme) sont effectivement appliquées. Dans le cas de règles concurrentes, toutes les possibilités sont générées. Pour choisir parmi ces possibilités la traduction la plus probable, nous avons utilisé une approche simple basée sur les modèles de langue (ML). Les ML sont largement utilisés pour la traduction artificielle, la transcription de l'oral ou encore la recherche d'information (Charniak, 1993). Cependant, dans ces cadres, les ML sont utilisés pour associer une probabilité à une séquence de mots, alors que dans notre cas, la probabilité va au contraire être associée à un mot considéré comme une séquence de lettres. Plus formellement, avec les notations standard on a :

$$P(w) = \prod_{i=1}^{m} P(l_i|l_1, ..., l_{i-1}) \quad \text{pour un terme } w \text{ composé des lettres } l_1, l_2, ..., l_m$$

En pratique, les probabilités $P(l_i|...)$ sont estimées à partir des termes biomédicaux de la langue cible, décomposés en n-grammes de lettres, issus de la liste d'exemples. Pour prévenir le problème des séquences de lettres non vues, les probabilités sont en réalité calculées à partir d'un historique réduit aux $n-1$ lettres précédentes (*i.e.* $P(l_i|l_{i-n+1}, ..., l_{i-1})$) et un lissage simple est appliqué. Dans les expériences présentées ci-après, n est fixé à 7 lettres.

Intuitivement, le ML va favoriser les traductions qui *ressemblent* à des termes biomédicaux bien formés dans la langue cible. Parmi toutes les traductions proposées pour un terme de la langue source, on conserve donc finalement celle qui obtient la probabilité la plus forte selon le ML appris. Par ailleurs, il est intéressant de noter qu'en plus de choisir la traduction la plus probable, cette technique nous donne un facteur de confiance sur la traduction retenue.

3.3 Commentaires sur la technique de traduction

Deux points concernant la technique de traduction proposée méritent d'être mentionnés. Il est tout d'abord intéressant de constater que l'approche que nous avons adoptée peut être rapprochée du cadre usuel de la traduction artificielle statistique dans lequel le but est de traduire une séquence de mots f dans une langue source en une séquence e dans la langue cible en cherchant e maximisant $P(e) \cdot P(f|e)$ (Brown *et al.*, 1993). Le terme $P(f|e)$ représente la probabilité que la séquence f soit la traduction de e. Ces probabilités, qui sont estimées à partir d'un corpus aligné, peuvent être rapprochées de nos règles de réécriture et leur score. Le terme $P(e)$ sert à vérifier que la séquence proposée soit bien formée ; son fonctionnement est en tout point similaire au modèle de langue que nous utilisons, si ce n'est que nos séquences sont composées de lettres et non de mots. Il y a cependant quelques différences importantes avec notre approche, principalement induites par la nature de nos données. Ainsi, dans le cadre de la traduction artificielle de textes, les différences d'ordonnancement des mots entre la langue source et la langue cible sont des problèmes difficiles à résoudre et mènent à construire des modèles de traduction compliqués. Dans notre cas, ce problème est quasiment inexistant : l'ordre des morphes, et donc des lettres composant les termes, varie peu d'une langue à l'autre. Le fait de manipuler des lettres et non des mots nous permet aussi d'utiliser un ML avec un historique de taille importante sans craindre de tomber trop souvent sur des séquences non observées ; les combinaisons possibles de lettres sont en effet bien moins nombreuses que les combinaisons de mots.

Le deuxième point à noter concerne une des limites évoquées par Claveau & Zweigenbaum (2005b) à propos de leur technique de traduction de termes. Ces derniers indiquent en effet que leur approche par transducteur ne peut pas prendre en compte des informations sur les termes comme les parties-du-discours (PoS). Cela a pour effet de générer des erreurs et de complexifier l'apprentissage dans le cas de mots polyfonctionnels comme *linguistique* qui se traduira différemment selon qu'il soit nom ou adjectif. Dans notre cas, cette limite est levée puisque l'ajout de ces informations se fait très naturellement avec le modèle de langues. Les probabilités peuvent en effet être calculées en incluant la partie-du-discours de la séquence en cours de traitement (on calcule alors les scores en fonction des $P(l_i|l_{i-n+1}, ..., l_{i-1}, PoS)$).

4 Évaluation de la traduction de terme

4.1 Description des données

Deux jeux de données sont utilisés pour nos expériences de traduction. Le premier est une collection de termes français-anglais issue du dictionnaire médical Masson (`http://www.atmedica.com`). C'est la même collection que celle utilisée dans Claveau & Zweigenbaum (2005b), ce qui nous permettra de comparer les résultats. Seules les paires composées de termes simples dans la langue source et dans la langue cible, hors acronymes, sont conservées. La liste bilingue ainsi constituée contient environ 12 000 paires de termes.

Le second jeu de termes multilingues est le Métathésaurus de l'UMLS (Bodenreider, 2004). Cette collection de thésaurus rassemble des termes biomédicaux dans 17 langues et associe à chacun des termes un identifiant de concept indépendant des langues, le Concept Unique identifier, CUI. Ces CUI nous permettent donc de constituer des ensembles de paires de termes bilingues. Là encore, seuls les termes simples non acronymes sont conservés.

4.2 Méthode d'évaluation

Pour l'évaluation de notre technique, nous suivons une approche standard : la liste initiale de paires de termes est découpée en deux ensembles, le premier sert pour l'apprentissage (inférence de règles et modèle de langue), et le second, composé de 1 000 paires, sert de jeu de test. Une fois les règles et le modèle de langue appris sur le jeu d'entraînement, nous l'appliquons à chaque terme d'entrée du jeu de test. Nous comparons alors la traduction proposée avec celle attendue. Si les deux chaînes de caractères sont identiques, la traduction est considérée correcte ; dans tous les autres cas, elle est considérée incorrecte.

Les résultats sont évalués en terme de précision (pourcentage de traductions correctes générées). Cependant, puisque le modèle de langue nous fournit un indice de confiance, on peut décider de ne conserver que les traductions dont l'indice est supérieur à un certain seuil. Un seuil élevé doit favoriser la précision au détriment du nombre de traductions proposées, et vice-versa. À la manière de courbes rappel-précision, nous représentons donc ci-après la précision suivant le pourcentage de mots traduits pour tous les seuils possibles d'indice de confiance.

4.3 Résultats

4.3.1 Traduction entre le français et l'anglais

Pour cette première expérience, nous nous intéressons à la traduction entre le français et l'anglais. Comme précisé précédemment, nous utilisons le jeu de données Masson qui nous permet de comparer directement nos résultats à ceux de Claveau & Zweigenbaum (2005b; 2005a) dont nous reportons les résutats ci-après. Nous utilisons aussi les informations de parties-du-discours dans le modèle de langue. Les figures 2 et 3 présentent les graphes de précision des traductions générées sur les ensembles de test pour les deux sens de traduction. Dans des langues proches comme le sont le français et l'anglais, beaucoup de termes spécialisés sont identiques. Comme simple *baseline*, nous calculons donc la précision qu'obtiendrait un système proposant systématiquement un terme comme sa propre traduction ; cette précision minimale donne ainsi une idée de la difficulté de la tâche de traduction.

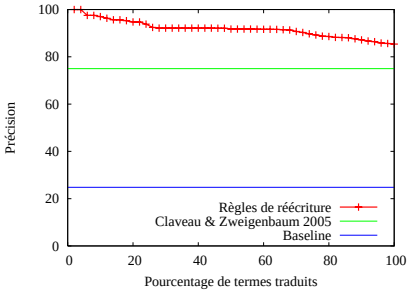


FIG. 2 – Performances de traduction du français vers l'anglais

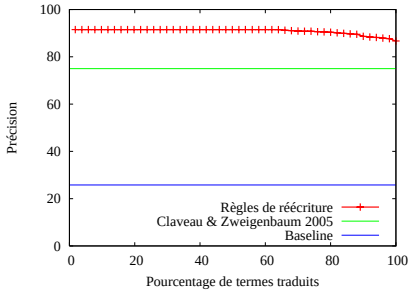


FIG. 3 – Performances de traduction de l'anglais vers français

Notre approche obtient de très bons résultats : 85.4 % de précision pour 100 % des mots traduits

pour le sens français-anglais et 84.8 % pour le sens inverse. Dans les deux cas, l'amélioration par rapport à l'approche par transducteur (Claveau & Zweigenbaum, 2005b) est de 10 %. Cela s'explique principalement par la souplesse de notre approche par règles qui lève la contrainte de déterminisme de l'approche par transducteur imposant qu'une séquence ne puisse se traduire que d'une façon, limitant et complexifiant ainsi l'apprentissage en présence de données bruitées, d'exceptions ou de mots polyfonctionnels. Concernant l'utilisation des modèles de langue, on remarque qu'utiliser les ML sans prendre en compte des parties-du-discours fournit des précisions légèrement plus faibles (82.6 % pour le sens français-anglais et 84.8 % pour l'autre sens). Et choisir le candidat au hasard parmi les différentes traductions génerées plutôt qu'utiliser les scores de ML mène à une précision d'environ 50 %. Ces deux résulats montrent bien l'intérêt des ML pour choisir la meilleure traduction et le bien-fondé de l'inclusion des parties-du-discours dans ces ML. Enfin, notre technique est très largement au-dessus de la *baseline*, mais il convient de noter que celle-ci montre que 25 % des termes biomédicaux sont identiques en français et en anglais, ce qui semble indiquer que les deux langues sont suffisamment proches pour rendre la tâche d'apprentissage relativement aisée.

4.3.2 Autres paires de langues

Nous répétons les expériences avec d'autres paires de langues disponibles dans l'UMLS. Parmi les différentes combinaisons de langues possibles, nous n'en présentons ci-dessous que quelques unes. La figure 4 présente les résultats obtenus avec deux langues réputées proches : l'espagnol et le portugais. Les résultats sont très bons : 87.9 % des termes portugais sont correctement traduits en espagnol quand on traduit tous les termes (*i.e.* quand aucun seuil pour le ML n'est fixé) ; dans le sens inverse, ce sont 85 % des termes qui sont correctement traduits. Ces bons résultats ne sont pas surprenants : l'espagnol et le portugais sont très proches, et comme le soulignent les très hautes *baselines*, beaucoup de mots sont en fait identiques.

Nous présentons maintenant les résultats obtenus par la traduction de diverses langues vers l'anglais — cas le plus à même d'être utilisé en pratique. La figure 5 présente les performances de la traduction de l'espagnol et du portugais vers l'anglais. Les résultats sont plutôt bons : 71.7 % des termes espagnols et 71.5 % des termes portugais sont correctement traduits quand toutes les traductions sont gardées (*i.e.* aucun seuil de ML n'est fixé). Ces résultats sont conformes avec la proximité de l'espagnol et du portugais illustrée dans l'expérience précédente. La traduction de l'italien et du tchèque vers l'anglais (figure 6) donne également des résultats comparables : au pire cas, 70 % des termes italiens et 75.5 % des termes tchèques sont correctement traduits.

Comme nous l'avons dit précédemment, notre technique de traduction peut traiter des langues avec des alphabets différents, pourvu qu'elles montrent des régularités pouvant être apprises automatiquement. Pour illustrer cela, nous nous intéressons à la paire russe-anglais. La figure 7 présente les résultats obtenus ; la précision minimale obtenue ici est de 57.5 % (du fait de la différence d'alphabet, la *baseline* est ici à 0). Bien qu'inférieurs aux autres paires de langues, ces résultats sont relativement bons étant donnée la difficulté apparente de la tâche. Cela met en exergue l'emploi en russe des mêmes racines gréco-latines que pour les autres langues étudiées.

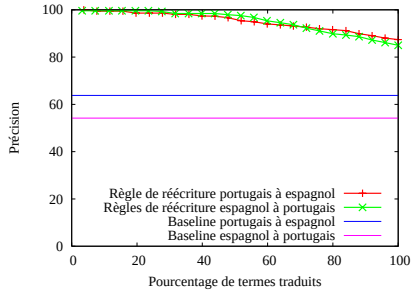


FIG. 4 – Performances de traduction espagnol - portugais

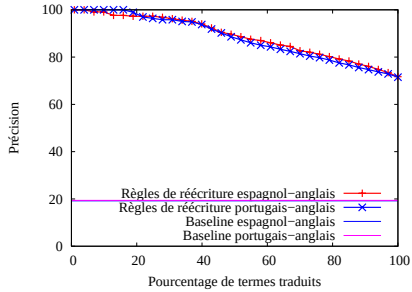


FIG. 5 – Performances de traduction portugais/espagnol vers anglais

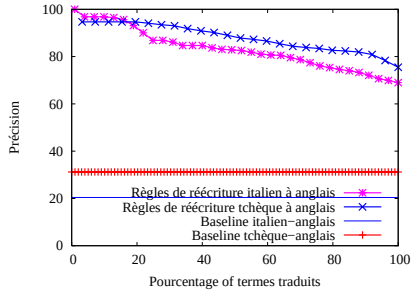


FIG. 6 – Performances de traduction de l'italien/tchèque vers l'anglais

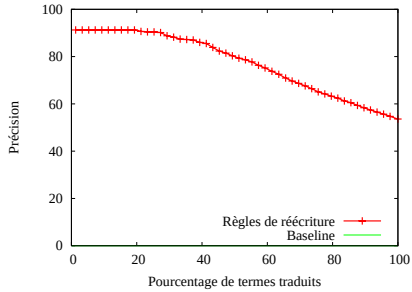


FIG. 7 – Performances de traduction du russe vers l'anglais

5 Conclusion

Notre technique de traduction à base de règles de réécriture capture automatiquement les régularités existant dans le domaine biomédical entre les termes des langues sources et cibles. De ce fait et sans surprise, la principale cause des traductions erronées observées est l'absence de lien morphologique entre le terme source et le terme cible. C'est évidemment souvent le cas pour la paire russe-anglais, mais aussi pour des paires de langues pourtant proches (*e.g. asimiento/grip* pour espagnol-portugais ou *embrochage/pinning* pour français-anglais). Les évaluations menées montrent cependant que ces cas sont assez rares pour que notre technique offre des taux de précision variables selon les langues mais très bons dans l'ensemble, et supérieurs à ceux disponibles dans l'état de l'art. De plus, l'utilisation de modèles de langues pour donner un score à chaque traduction permet de contrôler la précision souhaitée en fixant ou non un seuil à dépasser.

Parmi les perspectives ouvertes par ce travail, la traduction des termes complexes (composés de plusieurs mots, comme *col du fémur*) est l'une des plus importantes pour assurer une bonne couverture des terminologies à traduire (ils représentent par exemple environ 50 % de la terminologie MeSH). Enfin, dans un cadre plus applicatif, l'utilisation de cette technique dans un cadre de recherche d'information translingue (traduction de requêtes de la base PubMed) est à l'étude et donne des premiers résultats encourageants (Claveau, 2007).

Références

AHRENBERG L., ANDERSSON M. & MERKEL M. (2000). *A knowledge-lite approach to word alignment*, chapter 5, p. 97–138. In (Véronis, 2000).

BODENREIDER O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, **32**(D267-D270).

BROWN P. F., PIETRA V. J. D., PIETRA S. A. D. & MERCER R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, **19**(2).

CHARNIAK E. (1993). *Statistical Language Learning*. Cambridge, Massachusetts: MIT Press.

CLAVEAU V. (2007). Traduction automatique de termes biomédicaux pour la recherche d'information interlingue. In *Actes de la Conférence en Recherche d'Information et Applications, CORIA'07*, St-Étienne, France.

CLAVEAU V. & ZWEIGENBAUM P. (2005a). Automatic translation of biomedical terms by supervised transducer inference. In *Proceedings of the 10th Conference on Artificial Intelligence in Medicine, AIME 05*, Lecture Notes of Computer Science, Aberdeen, Écosse: Springer.

CLAVEAU V. & ZWEIGENBAUM P. (2005b). Traduction de termes biomédicaux par inférence de transducteurs. In *Actes de la conférence Traitement automatique des langues naturelles, TALN'05*, Dourdan, France.

FLUHR C., BISSON F. & ELKATEB F. (2000). *Parallel text alignment using crosslingual information retrieval techniques*, chapter 9. In (Véronis, 2000).

FUNG P. & MCKEOWN K. (1997). A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation*, **12**(1/2), 53–87.

GALE W. & CHURCH K. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the 4th Darpa Workshop on Speech and Natural Language*, p. 152–157, Pacific Grove, CA, États-Unis.

KNIGHT K. & GRAEHL J. (1998). Machine transliteration. *Computational Linguistics*, **24**(4), 599–612.

NAMER F. (2005). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. In *Actes de la conférence Traitement Automatique des langues naturelles, TALN'05*, Dourdan, France.

SCHULZ S., MARKÓ K., SBRISSIA E., NOHAMA P. & HAHN U. (2004). Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, p. 813–819, Genève, Suisse.

TSUJI K., DAILLE B. & KAGEURA K. (2002). Extracting french-japanese word pairs from bilingual corpora based on transliteration rules. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC'02*, p. 499–502, Las Palmas de Gran Canaria, Espagne.

J. VÉRONIS, Ed. (2000). *Parallel Text Processing*. Dordrecht: Kluwer Academic Publishers.

WAGNER R. A. & FISCHER M. J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery*, **21**(1), 168–173.

Session

Outils

TiLT correcteur de SMS : évaluation et bilan qualitatif

Émilie GUIMIER DE NEEF, Arnaud DEBEURME, Jungyeul PARK
FTR&D/TECH/EASY – France Telecom R&D
2, avenue Pierre Marzin, 22300 Lannion Cedex, France
{emilie.guimierdeneef,arnaud.debeurme,jungyeul.park}
@orange-ftgroup.com

Résumé. Nous présentons le logiciel TiLT pour la correction des SMS et évaluons ses performances sur le corpus de SMS du DELIC. L'évaluation utilise la distance de Jaccard et la mesure BLEU. La présentation des résultats est suivie d'une analyse qualitative du système et de ses limites.

Abstract. This paper presents TiLT system which allows us to correct spelling errors in SMS messages to standard French. We perform Jaccard and Bleu metrics for its evaluation using the DELIC SMS corpus as a reference. We discuss qualitative analyses of system and its limits.

Mots-clés : SMS, SMS corpus, correction orthographique, TiLT, evaluation.

Keywords: SMS, SMS corpus, spelling correction, TiLT, evaluation.

1 Introduction

Les nouvelles formes de communication écrite (blogs, SMS, chats etc.) se caractérisent par de nombreux écarts vis-à-vis des conventions orthographiques standards. Ces écarts recensés par Anis (2002), confirmés par des études de corpus réels Bove (2005) et très brièvement rappelés ci-dessous, offrent un nouveau défi aux outils de correction automatique. En effet, comme l'ont montré Véronis et Guimier De Neef (2006), un simple recensement de couples graphie SMS / graphie standard ne suffit pas à répondre à la productivité et à la combinatoire des différents procédés d'écriture.

- Graphies phonétisantes et rébus : *g ht du kfé a+ (j'ai acheté du café à plus)*
- Abréviations diverses : *slt k f tu (salut que fais-tu ?)*
- Étirements graphiques : *ssuuuppperr ! hhhhhuuuuummmm !*
- Agglutinations : *g ésayé 2tapelé pl1 2foi (j'ai essayé de t'appeler plein de fois)*

Différentes motivations peuvent justifier la correction, ou plutôt la normalisation, de l'écriture SMS : l'extraction d'information, l'indexation, l'analyse de blogs, de wikis etc. La vocalisation des SMS destinés aux téléphones fixes a été l'occasion pour FTR&D d'adapter son logiciel TiLT à ce contexte. L'architecture globale du service est schématisée Figure 1.

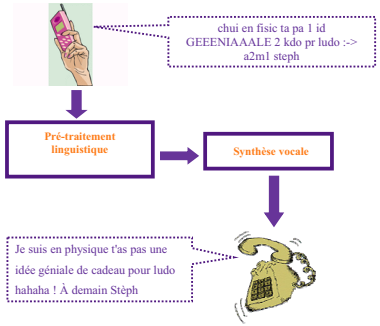


Figure 1 : correction des SMS avant vocalisation

Dans cet article, nous présentons une évaluation du correcteur de SMS TiLT en mesurant ses performances sur le corpus de SMS du laboratoire DELIC. Dans une première section, nous décrivons le logiciel et ses adaptations au contexte SMS. Nous présentons le corpus du DELIC en seconde section. Les mesures utilisées et les résultats de l'évaluation sont fournis en troisième section ainsi qu'une analyse qualitative des résultats obtenus. Nous terminons l'article par différentes perspectives de recherche.

2 TiLT correcteur de SMS

Le logiciel TiLT est une solution d'analyse linguistique modulaire permettant différents types de traitements automatiques comme la correction orthographique, le pré-traitement linguistique de documents avant indexation, l'extraction d'informations, la traduction automatique etc. Nous présentons ci-dessous l'architecture et les particularités de la solution TiLT pour la correction orthographique dont l'application SMS est une instance particulière.

Le logiciel fait intervenir séquentiellement trois briques : (i) un module de segmentation, (ii) un module d'analyse lexicale avec ou sans méthodes correctives, (iii) un module d'analyse en chunking. Des données linguistiques spécifiques au contexte SMS ont été développées à partir d'observations linguistiques et de tests sur corpus. L'ensemble des données symboliques utilisées par le logiciel est développé à partir d'une expertise humaine.

2.1 Segmentation

La segmentation permet le découpage et le typage des différents segments du message en entrée. Le typage des segments permet de différencier ensuite le traitement qu'il convient de

leur associer : seuls les segments de type MOT sont envoyés à l'analyse lexicale. La description des segments se fait au moyen d'expressions régulières compilées par FLEX[1].

Pour le traitement des SMS, une modification importante des données de segmentation est rendue nécessaire principalement pour l'identification des smileys et pour l'inclusion massive des chiffres et des symboles dans les mots.

```
Phrase analysée : tu vi1 2m1 a+ ;-)

0  1 : "tu"  : MOT
1  2 : "vi1" : MOT_AVECCHIFFRE
2  3 : "2m1" : MOT_AVECCHIFFRE
3  4 : "a+"  : MOT_AVECSYMB
4  5 : ";-)" : SMILEY
```

Figure 2 : segmentation TiLT

2.2 Lexique

Le lexique du français utilisé par TiLT comporte environ 100 000 entrées incluant une base de mots-composés. Pour le traitement des SMS, ce lexique a été enrichi d'une base d'un millier d'abréviations recueillies sur le web ou compilées d'après relevés sur corpus. Parmi ces abréviations, on recense des sigles (*atd = à ta disposition*, *tvb = tout va bien*…), des squelettes consonantiques (*slt a vs ts = salut à vous tous*), des troncations (*adr = adresse*) etc. Le lexique SMS inclut également une base de prénoms de 3 000 entrées environ.

Un apprentissage sur corpus a permis de délimiter la liste des mots les plus fréquemment utilisés dans un contexte SMS afin de favoriser ces formes par rapport aux autres mots de la base lexicale.

2.3 Les méthodes correctives

Au cours de l'analyse lexicale, le lexique est consulté et renvoie l'ensemble des informations disponibles : lemme, orthographe standard, catégorie grammaticale, traits morpho-syntaxiques. Cette consultation se fait selon différents modes dont des modes correctifs, certains ayant été spécialement adaptés au contexte SMS.

1. **Correction phonétique** : le module de correction phonétique, basé sur un transducteur appris par alignement des formes phonétiques et graphiques des mots d'une langue, a été enrichi de règles permettant la phonétisation des symboles, chiffres et lettres utilisés pour leur valeur phonétique en écriture SMS.

2. **Correction par découpage morpho-syntaxique** : des observations sur corpus (Bove 2005) ont montré que l'agglutination de mots non généralisée n'intervient pas au hasard mais concerne de façon privilégiée les séquences avec clitiques (*jtrappelle*, *gspère qtu va bien*…), avec préposition (*g ésayé 2tapelé pl1 2foi*), les séquences déterminant/nom (*c le foot ki te mé ds 7éta?*) ou les formes lexicales complexes

[1] http://fr.wikipedia.org/wiki/Flex_%28GNU%29.

(*Keske tu deviens?*) etc. Le module de correction par découpage a été associé à des données spécifiques pour permettre l'expansion de ces formes compactées.

3. **Tolérance à la répétition** : l'une des particularités de l'écriture SMS est la présence de marques expressives dans l'écriture. Ces marques incluent les smileys, des jeux sur les signes de ponctuation (*on est sur la plage!!!!!!!!!!*), l'utilisation de la casse (*Encore un grand MERCI à tous les deux*) mais aussi l'étirement graphique (*c la foliiiiiiiie !!*). Une méthode corrective dite de tolérance à la répétition a été spécialement développée pour restituer les orthographes non étirées.

Ces méthodes calculent dynamiquement les corrections possibles dans les SMS, ce qui répond à la créativité orthographique des utilisateurs de SMS.

2.4 La grammaire de chunking

L'analyse en chunking pratiquée par TiLT utilise une grammaire hors contexte de 2 000 règles environ. Son rôle est de permettre de choisir la correction adaptée pour un mot étant donné son contexte syntaxique local.

En SMS, deux particularités augmentent la difficulté du chunking :

1. l'ambiguïté en général et celle des mots outils en particulier, comme dans le paradigme suivant où le caractère *c* se normalise de 4 façons différentes :

Voilà c fini ca c bin passé	=>	*voilà c'est fini ça s'est bien passé*
Je c pa ki c	=>	*je sais pas qui c'est*
Jespere k vou descendé c soir	=>	*j'espère que vous descendez ce soir*

2. la ponctuation souvent absente qui cesse de jouer son rôle de césure :

alor t soulagé moi la jaten 1h ca me soule j menui a mourir biz
G un empechement previen sophie gros poutou à demain

Pour contrer ces difficultés, la grammaire a été enrichie de règles prenant en compte des structures particulièrement fréquentes en contexte SMS (interrogatives, formes présentatives introduites par *c'est*..., constructions modales etc.). En particulier, une grammaire locale des formules de politesse, s'appuyant sur des déclencheurs tels que *salut*, *bisous* etc. s'est avérée indispensable pour aider à la détection des noms propres (*HELLO RÉJANE, A BIENTO JOHAN, coucou Nénette, a+ Reno* etc.)

3 Le corpus de SMS du DELIC

3.1 Présentation du corpus

Dans le cas d'un contrat collaboratif avec France Télécom R&D, le laboratoire DELIC de l'université d'Aix en Provence a collecté, avec le concours de ses étudiants, un corpus d'environ 9 700 messages SMS correspondant à un peu plus de 132 000 mots. Ces messages ont été corrigés semi-automatiquement puis révisés manuellement pour constituer une base alignée de SMS et de transcriptions. Le Tableau 1 fournit un extrait de ce corpus.

tu pe tokup du cha 2m1?	Tu peux t'occuper du chat demain ?
Je sui dvt ché toi	Je sui devant chez toi.
Jarriv!!tnev pa!!	J'arrive ! [ne] t'énerve pas !
Je conte sur toi 2m1 a la danse!	Je compte sur toi demain à la danse
Dsl pour ier..enormes bisous	Désolé(e) pour hier… énormes bisous.
pti coucou d vacs!! C terrible c bo !! je vs racontré	Petit coucou des vacances ! C'est terrible, c'est beau ! Je vous raconterai !

Tableau 1 : Extrait du corpus aligné SMS / français standard du DELIC

Une description fine des conventions de correction utilisées dans la transcription du corpus peut être trouvée dans Hocq (2006). Globalement, les orthographes phonétiques, erronées, abrégées sont corrigées. Il en va de même des marques de ponctuation qui sont ramenées à l'usage standard. Des alternances d'accord sont proposées en cas d'ambiguïté (*dsl = désolé ou désolée*). Certains mots manquants, *ne* négatif, pronom sujet principalement, sont notés entre crochets. Les sigles et abréviations sont étendus (*mdr = mort de rire, dvt = devant* etc.). Des données personnelles ont été rendues anonymes. Les numéros de téléphone ont été remplacés par 01 02 03 04 05. La balise <NOM> se substitue aux noms de personne identifiables : Michel Dupont est remplacé par <NOM>. Par contre, les prénoms isolés ont été conservés.

Ce corpus peut être caractérisé par quelques chiffres. La taille moyenne des messages SMS est de 14,5 mots. Le nombre moyen de caractères par message est de 66,7 caractères[2]. Le rapport entre le nombre de caractères utilisés dans la correction et le nombre de caractères présents dans le SMS source, nous donne un taux de compression de l'écriture SMS de l'ordre de 20%.

3.2 Post-traitements pour l'évaluation

La correction fournie par TiLT ne suivant pas toujours les mêmes normes que la correction manuelle proposée dans le corpus du DELIC, une phase de normalisation des sorties de TiLT et d'enrichissement du corpus du DELIC a été effectuée. Les principales divergences portent sur la normalisation des heures (10h30 vs 10 h 30), des nombres notés en lettres ou en chiffres, des unités de mesure (km vs kilomètre), de certaines abréviations conservées ou étendues dans les corrections automatiques ou dans le corpus du DELIC. Certains choix de restitution TiLT en rapport avec l'application de vocalisation ont été revus : la restitution des smileys sous forme de balise par exemple (le smiley ;-) devient <SMILEY>).

Après post-traitement du corpus du DELIC, à chaque SMS correspond en moyenne 1,2 transcription standardisée. C'est cette version post-traitée du corpus qui sert de référence pour notre évaluation. Le Tableau 2 montre un extrait du corpus après les post-traitements.

[2] Le nombre maximum de caractères autorisés pour la saisie d'un SMS sur téléphone mobile est de 160 caractères. Certains téléphones autorisent la saisie de messages plus longs qui sont alors découpés avant d'être envoyés au destinataire.

tu t es planté en math t a eu 1 sal note	Tu t'es planté en math, t'as eu une sale note. Tu t'es planté en math, tu as eu une sale note.
tu pe venir me prendre a aix ver 2h30? merci ta couzine bizz	Tu peux venir me prendre à Aix vers 2 h 30 ? Merci. Ta cousine. Bise. Tu peux venir me prendre à Aix vers 2h30 ? Merci. Ta cousine. Bise. Tu peux venir me prendre à Aix vers 2 h 30 ? Merci. Ta cousine. Bisou. Tu peux venir me prendre à Aix vers 2h30 ? Merci. Ta cousine. Bisou.
TU PE VENIR	Tu peux venir.

Tableau 2 : Extrait du corpus aligné SMS / français standard après post-traitements

4 Évaluation

Les objectifs de cette évaluation sont doubles. Il s'agit d'une part de trouver des indicateurs permettant de quantifier objectivement les performances pour pouvoir en suivre les évolutions dans le temps. Il s'agit également de repérer les phénomènes de l'écriture SMS résistants pour mieux cerner les limites de notre approche et prévoir des extensions. En conséquence, la partie évaluation objective est suivie d'une analyse qualitative des résultats obtenus.

4.1 Évaluation objective

Parmi les mesures fréquemment utilisées pour mesurer les performances des traducteurs statistiques ou des systèmes de reconnaissance vocale, nous en avons retenu deux pour l'évaluation de notre correcteur de SMS : la mesure BLEU (Papineni et al. 2002) et le coefficient de Jaccard. La prise en compte ou pas de l'ordre des mots distingue les deux types de mesure. Le coefficient de Jaccard[3] considère la phrase comme un sac de mots, tandis que la mesure BLEU[4] prend en compte les n-grams et pénalise les corrections qui divergent quant à l'ordre des mots.

---

[3] Coefficient de Jaccard = $\frac{\text{nb mots dans l'intersection}}{\text{nb mots dans l'union}}$

nb mots dans l'intersection = nombre de mots communs entre la solution et la référence la plus proche.

nb mots dans l'union = nombre de mots de la solution + nombre de mots de la référence – nombre de mots communs. S'il y a un seul mot en commun, cette mesure n'est pas nulle.

[4] $BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{\log(\text{nb occurrences n - gram / nb n - grams})}{N}\right)$

nb occurrences n-gram = nombre de n-gram commun avec au moins une référence.

nb n-gram = nombre de n-gram dans la phrase à évaluer = taille de la phrase – (n – 1)

BP = Brevity Penalty = min (1, exp(1 – Min Nb Mots Reférence / Nb Mots Solution))

Si la solution a au moins le même nombre de mots que la plus petite des références, BP = 1 (pas de pénalité).

Pour la mesure BLEU standard N = 4. La mesure BLEU est donc une moyenne logarithmique (=géométrique) entre les taux de 1-gram, 2-gram, etc, en commun avec les références. Cette moyenne est pondérée par un facteur entre 0 et 1 : Brevity Penalty. Plus il y a de références, meilleur sera le score. En mesure BLEU standard

La solution de correction TiLT, à la différence d'un système de traduction automatique, ne reformule pas le message SMS. Elle corrige les mots dans l'ordre dans lequel ils sont formulés. Les risques d'erreurs liées à l'ordre des mots sont donc assez faibles. Néanmoins, l'écriture SMS présentant de nombreux cas d'agglutinations à étendre, l'utilisation de la métrique BLEU nous a semblé appropriée.

En standard, la métrique BLEU prend en compte les n-grams jusqu'au 4-grams. Une seule erreur de correction située au milieu d'un message SMS bref, dont la taille est de 5 ou 6 mots, est très fortement sanctionnée par BLEU. Le tableau suivant montre quelques exemples de ce type, fortement pénalisés par BLEU alors que plutôt bien notés par Jaccard.

SMS source	Correction TiLT	Correction DELIC	Jaccard	BLEU
CT koi ldebu du mess?	c'était quoi début du message ?	C'était quoi le début du message ?	0,85	0
COMEN FAI T ON ALOR?	COMMENT fait-t-ON alors ?	Comment fait-on alors ?	0,8	0
Lé fete toute seul c cool	Les fêtes toute seul c'est cool	Les fêtes toute seule, c'est cool	0,75	0

Tableau 3 : Comparaison de résultats obtenus avec les métriques Jaccard et avec BLEU

Considérant BLEU comme peu informatif sur les messages brefs très fréquents en SMS, un paramètre a été ajouté au calcul de la mesure BLEU permettant de faire varier *n* dans le calcul des n-grams en fonction de la taille du message. Pour un message de 4 ou 5 mots, BLEU s'arrêtera aux bi-grams, pour un message de 6 ou 7 mots, on s'arrête aux tri-grams etc. Avec ce paramètre, la mesure BLEU fait d'avantage sens pour les corpus SMS et la cohérence entre Jaccard et BLEU est plus importante.

Le Tableau 4 donne les résultats obtenus sur les 9 575 SMS du corpus DELIC avec ces trois mesures. Précisons que la casse et les signes de ponctuation ont été ignorés pour le calcul. En cas de référence multiple pour un SMS (cf. Tableau 2), le score retenu est le meilleur score obtenu sur l'ensemble des références possibles.

Jaccard	BLEU standard	BLEU pondéré
0,769	0,681	0,712

Tableau 4 : Scores BLEU et Jaccard obtenus sur le corpus SMS du DELIC

4.2 Évaluation qualitative

Une étude des scores obtenus SMS par SMS permet de voir qu'environ 25% du corpus reçoit le score maximal de 1. Parmi les SMS source, on trouve des exemples parfaitement ou quasi-parfaitement orthographiés et non dégradés par la correction automatique :

(N = 4), si aucun tétragramme est commun avec une référence, le score est 0. Si la solution est courte, la moindre différence avec les références donnera un score de 0. Il est donc préférable d'ajuster le paramètre N en fonction de la taille de la solution. N = min (4 , Nb Mots Solution / 2) par exemple.

Soliel et piscine tout va bien bisous
Sommes à arles maman
Souper à la maison ce soir pour feter le début de mes vacances

mais également des exemples à l'écriture typiquement SMS très bien corrigés :

ojrd8 jv ala pi6n tu ve vnir?	Aujourd'hui je vais à la piscine, tu veux venir ?
noubli pa ke jseré tjs la pr toi	N'oublie pas que je serai toujours là pour toi
bjr, vs avé le tps pr 1 kf ? avt 15h.G du taf	Bonjour, vous avez le temps pour un café ? Avant 15h. J'ai du taf.

Tableau 5 : SMS dont la correction reçoit un score de 1

L'analyse des mauvais scores montre trois grands types de limites dans notre système actuel. La première tourne autour de la primitive "mot" prise comme point de départ pour effectuer ses hypothèses correctives. On remarque, en effet, que les plus mauvais scores sont obtenus sur les SMS présentant une absence de séparateur généralisée ou avec un séparateur peu classique de type casse ou symbole particulier. Ce type de phénomène est présent dans 1 à 2% des messages SMS.

SMS	**Correction manuelle**	**Correction TiLT**
TuTcouchéto!Cbi1! moijaVpEr2pareusir adormir.onsphon...	Tu t'es couché tôt ! C'est bien ! Moi j'avais peur de pas réussir à dormir. On se phone...	TuTcouchéto ! c'est bien ! MoijaVpEr2pareusir dormir . Onspho
B o n n e f e t e p r o f i t e s b i e n d e v o t r e d e r n i e r j o u r d e v a c a n c e	Bonne fête, profite bien de votre dernier jour de vacances.	Bonnefeteprof il t'est biendevotrede rnierjourdeva cance

Tableau 6 : SMS présentant une absence de séparateur

De même, sur les méthodes correctives, notre approche trouve ses limites quand un même segment cumule différents procédés d'écriture : phonétique et agglutination (*je ne pep a mpaC dtoi => je ne peux pas me passer de toi*), étirement et phonétique : (*G haaaaaaateuh => j'ai hâte)* etc.

La solution à ces deux difficultés est sans doute à aller chercher du côté des méthodes employées en reconnaissance de la parole pour segmenter le signal acoustique. Le problème de l'absence de séparateur n'est pas sans rappeler une langue comme le chinois pour laquelle des algorithmes de segmentation ont été développés.

La seconde limite vient de l'absence d'apprentissage dans le développement des données. L'indisponibilité d'un corpus aligné lors de la mise au point de la solution en est la cause. On peut espérer pouvoir maintenant tenter des expériences pour extraire les n-grams fréquents et enrichir les lexiques d'expressions récurrentes telles que *rien de spécial*, etc. Apprendre des séquences fréquentes de mots pour pouvoir affiner les scores attribués aux différentes hypothèses etc.

Cet apprentissage permettrait sans doute d'ouvrir l'espace des corrections aux mots dont l'orthographe est connue du lexique[5]. En effet, la stratégie utilisée actuellement a l'avantage de ne pas dégrader les messages bien écrits mais a l'inconvénient de laisser de nombreux homophones hétérographes non corrigés :

SMS	**Correction manuelle**	**Correction TiLT**
Tu me racontera dis	Tu me raconteras, dis	Tu me racontera dis
ns avons dormis	Nous avons dormi	Nous avons dormis
T ou au moi daout?	T'es où au mois d'août	T'es où au moi d'août

Tableau 7 : homophones hétérographes non corrigés

La troisième lacune concerne la trop grande localité des règles de grammaire. Une grammaire vérifiant les principales contraintes de sous-catégorisation aiderait à orienter certains choix de correction. Dans l'exemple *jme langui tro dy aller*, TiLT échoue à corriger faute de lien entre *languir* et son dépendant *d'y aller*. De même, l'accord sujet verbe n'est pas vérifié dans *tes vacs se pass bi1* qui est corrigé en *tes vacances se passe bien*. L'utilisation d'une grammaire vérifiant des contraintes entre tête et dépendant serait à expérimenter.

Enfin, des travaux autour des noms propres restent également à faire même si étant donné leur anonymisation dans ce corpus, il est difficile de tirer des conclusions définitives sur ce point. Néanmoins, on remarque sans surprise que l'identification des prénoms sans contexte déclencheur connu des lexiques est générateur de nombreuses erreurs : *Julie c jb ap moi => Julie ce gibet après moi*, *gros bisous à vous tous caro => gros bisous à vous tous carreau*.

5 Bilan et perspectives

Les mesures BLEU et Jaccard pratiquées pour l'évaluation des performances de TiLT correcteur de SMS montrent que la solution est efficace à 75% environ. L'adéquation des mesures utilisées reste bien entendu à discuter. En particulier, il pourrait être intéressant de pondérer les scores obtenus par la difficulté a priori de la correction ; cette difficulté étant quantifiable par une distance entre le SMS et sa transcription. Il faut noter également que pour une application de correction de SMS avant synthèse vocale, il faudrait prévoir une évaluation par l'usage : les erreurs sur les homophones seraient sans doute moins pénalisantes.

La correction pratiquée par TiLT n'exploite pas de données apprises sur corpus. Une évaluation sur un autre corpus permettrait de s'assurer de la stabilité de la solution. C'est pourquoi, nous espérons pouvoir faire cette même évaluation sur le corpus de Fairon et al. (2006).

Les 25% à 30% de mauvaises corrections montrent les limites de l'approche corrective actuelle : pas de correction des homophones hétérographes, pas de segmentation des messages sans séparateur, pas de mode correctif hybride etc. autant de phénomènes qui nous montrent

[5] Sauf dans certains cas très fréquents d'erreur comme la confusion participe passé en *–é* et infinitif en *-er*.

la parenté entre les messages SMS et la langue orale et qui nous invitent à reconsidérer la place centrale de la notion de mot dans notre traitement.

Remerciements

Nous remercions Olivier Collin et Jean Véronis pour leurs conseils et Sabrina Hocq pour son travail de collecte et de transcription du corpus.

Références

ANIS, J., (1999). Chats et usages graphiques. *Internet, communication et langue française.* In Anis J. (éd.), Paris : Hermès, 71-90.

ANIS, J., (2001). *Parlez-vous texto ? Guide des nouveaux langages du réseau.*, Paris : Le cherche-midi éditeur.

ANIS, J., (2002). Communication électronique scripturale et formes langagières : chats et SMS., *Actes des journées « S'écrire avec les outils d'aujourd'hui »*, Université de Poitiers.

BOVE, R., (2005)., Étude de quelques problèmes de phonétisation dans un système de synthèse de la parole à partir de SMS, *Actes de RÉCITAL 2005*, Dourdan, 625-634.

FAIRON, C., KLEIN, J., PAUMIER, S., (2006)., *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête 'Faites don de vos SMS à la science'*, Presses universitaires de Louvain, Louvain-la-Neuve.

GUIMIER DE NEEF, É., VÉRONIS, J., (2004). 1 pw1 sr la kestion ;-), *Papier présenté à la Journée d'Étude de l'ATALA "Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.),* Paris.

HOCQ, S., (2006). Étude des SMS en français : constitution et exploitation d'un corpus aligné SMS – langue standard. *Rapport de Master II "Industries des Langues"*, Aix-en-Provence.

PAPINENI, K., ROUKOS, S., WARD, T., ZHU, W. J., (2002), BLEU: a method for automatic evaluation of machine translation, in *ACL-2002 : 40th Annual meeting of the Association for Computational Linguistics*, 311-318.

VÉRONIS, J., GUIMIER DE NEEF, É., (2006). Le traitement des nouvelles formes de communication écrite. In Sabah, G. (Éd.), *Compréhension automatique des langues et interaction*, 227-248, Paris: Hermès Science.

Vers un méta-EDL complet, puis un EDL universel pour la TAO

Hong-Thai NGUYEN[1], Christian BOITET[2]
GETALP, LIG
385, av. de la Bibliothèque, BP 53 F-38041 Grenoble cedex 9
{Hong-Thai.Nguyen,Christian.Boitet}@imag.fr

Résumé. Un "méta-EDL" (méta-Environnement de Développement Linguiciel) pour la TAO permet de piloter à distance un ou plusieurs EDL pour construire des systèmes de TAO hétérogènes. Partant de CASH, un méta-EDL dédié à Ariane-G5, et de WICALE 1.0, un premier méta-EDL générique mais aux fonctionnalités minimales, nous dégageons les problèmes liés à l'ajout de fonctionnalités riches comme l'édition et la navigation en local, et donnons une solution implémentée dans WICALE 2.0. Nous y intégrons maintenant une base lexicale pour les systèmes à « pivot lexical », comme UNL/U++. Un but à plus long terme est de passer d'un tel méta-EDL générique multifonctionnel à un EDL « universel », ce qui suppose la réingénierie des compilateurs et des moteurs des langages spécialisés pour la programmation linguistique (LSPL) supportés par les divers EDL.

Abstract. A "meta-EDL" (meta-Environment for Developing Lingware) for MT allows to pilot one or more distant EDL in order to build heterogeneous MT systems. Starting from CASH, a meta-EDL dedicated to Ariane-G5, and from WICALE 1.0, a first meta-EDL, generic but offering minimal functionalities, we study the problems arising when adding rich functionalities such as local editing and navigation, and give a solution implemented in WICALE 2.0. We are now integrating to it a lexical database for MT systems relying on a "lexical pivot", such as UNL/U++. A longer-term goal is to evolve from such a multifunctional generic meta-EDL to a "universal" EDL, which would imply the reengineering of the compilers and engines of the specialized languages (SLLPs) supported by the various EDLs.

Mot-clés : génie linguiciel, langages spécialisés pour la programmation linguistique, LSPL, environnement de développement, EDL, TAO, systèmes distribués hétérogènes.

Keywords: lingware engineering, specialized languages for linguistic programming, development environment, EDL, MT, heterogeneous distributed MT systems.

1 Introduction

Il existe des EDL (Environnements de Développement Linguistique) pour systèmes de TAO, plus ou moins complets. Ils sont tous construits autour d'une technologie spécifique. On peut citer Ariane-78 puis Ariane-G5 du GETA (Grenoble), Tapestry du CRDL (Singapour), ETAP-3 de l'IPPI (Moscou), et ceux des fournisseurs de système de TAO commerciaux, non disponibles pour la recherche.

Depuis 10 ans environ, on cherche à réaliser des systèmes de TAO hétérogènes, soit pour combiner plusieurs systèmes pour une nouvelle paire de langues (approche « multimoteur »

de Pangloss (Nirenburg and Frederking, 1994), VerbMobil (Ney H, Och & Vogel, 2000), etc.), soit pour construire un système de TAO fortement multilingue dont les composants peuvent être développés par différents groupes, avec des approches et des EDL différents, comme dans le projet UNL (Projet UNL).

Pour permettre le développement coopératif et distribué de ce type de système, une première étape consiste à développer un « méta-EDL » fonctionnant comme une interface avec plusieurs EDL distants, i.e. permettant d'éditer et de synchroniser les composants linguiciels (dictionnaires, grammaires, automates) et de combiner différents modules distants pour produire des traductions. Un problème intéressant est alors d'intégrer le plus possible de fonctions des EDL (navigation, aide à l'indexage des dictionnaires, etc.), sans devoir effectuer une réingénierie de ces EDL. Enfin, comme cette approche est par nature limitée, un but plus ambitieux est de construire un « EDL générique » pour le développement distribué de systèmes de TAO hétérogènes.

Cet article est organisé en trois parties. Nous détaillons d'abord les fonctionnalités des EDL et des méta-EDL, et en donnons une illustration avec CASH et WICALE 1.0. Dans la deuxième partie, nous montrons les problèmes posés par l'ajout à WICALE des fonctions d'édition et de navigation en local, ainsi que les solutions retenues pour leur implémentation dans WICALE 2.0. Dans la troisième partie, nous montrons la nécessité et la difficulté d'intégrer une base lexicale dans un méta-EDL, et décrivons PIVAX, une base lexicale multilingue organisée autour d'un « pivot lexical », en cours de construction. PIVAX pourra être utilisée non seulement pour la TAO, en particulier pour le projet UNL/U++, mais pour développer d'autres applications comme la recherche d'informations en contexte multilingue (CLIR).

2 EDL et méta-EDL pour la TAO

Un Environnement de Développement Linguistique (EDL) est un environnement de *programmation linguistique* qui connecte ou intègre un ou plusieurs *LSPL* (Langages Spécialisés pour la Programmation Linguistique). Un EDL permet aux développeurs linguistes de réaliser les opérations nécessaires (gestion, manipulation des linguiciels et des données, compilation, test, production) de façon transparente.

Un « méta EDL» permet de piloter à distance un ou des EDL distants. Pour l'instant, nous travaillons à la construction d'un méta-ED pour la TAO le plus puissant possible. Dans le futur, nous voulons développer un EDL « universel » permettant la réingénierie de tout EDL.

Les caractéristiques des EDL sont assez différentes de celles des IDE (environnement de développement de logiciels). Nous l'illustrerons avec deux exemples de méta-EDL existants.

2.1 Fonctionnalités d'un EDL

Table 1: comparaison des IDE et des EDL

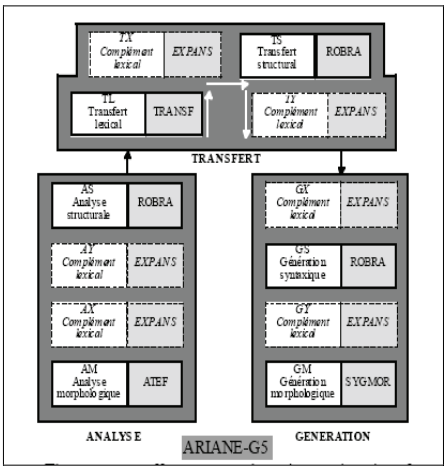
	IDE	EDL
Utilisateur	programmeur	linguiste, lexicographe, gestionnaire, utilisateur …
Type de composant	fonction, procédure, objet, module, paquetage…	variables, modèles, grammaire, automate, dictionnaire…
Taille de composant	petite (quelques pages)	grande (100K-1G entrées pour un gros dictionnaire de TA, 100-400 pages pour une grammaire d'analyse)
Type d'évolution	plutôt stable	en perpétuelle évolution

Un EDL de TAO complet doit offrir 4 « classes de fonctionnalités ».

- Préparation des composants linguiciels : (1) visualisation, (2) édition, (3) tri, (4) aide à l'indexage.
- Organisation en étapes et phases : (1) gestion de versions de test pour mise au point, (2) gestion de chaînes d'exécution (totale ou partielle), (3) génération de systèmes de TAO complet ou des parties de tels systèmes.
- Gestion de corpus d'essai : (1) création ou modification de corpus ou de textes, (2) passage de tests, (3) traduction de (parties de) corpus, (4) révision humaine.
- Actions globales : (1) extraction d'informations, (2) vérification de cohérence, (3) impressions avec filtrage et tri.

2.2 Les EDL de TAO existants

Le plus complet semble toujours être Ariane-G5 (Ariane-Y) du GETA (Figure 1), le seul qui permet de « créer » un nouveau système de TAO en quelques commandes, sans intervention d'informaticiens. Le processus de traduction se compose de trois étapes (analyse, transfert, génération), chaque étape étant composée de phases (obligatoires ou facultatives).



En analyse, on trouve :		
AM	analyse morphologique	**obligatoire**
AX	analyse expansive	facultative
AY	analyse expansive	facultative
AS	analyse structurale	**obligatoire**
En transfert, on trouve :		
TL	transfert lexical	**obligatoire**
TX	transfert expansif	facultative
TS	transfert structural	**obligatoire**
TY	transfert expansif	facultative
En génération, on trouve :		
GX	génération expansive	facultative
GS	génération syntaxique	**obligatoire**
GY	génération expansive	facultative
GM	génération morphologique	**obligatoire**

Figure 1 : étapes, phases et LSPL de l'EDL Ariane-G5

Pour générer un système de TAO, l'utilisateur (un développeur linguiste) écrit un linguiciel pour chaque phase avec le LSPL adéquat, le compile et génère une « chaîne d'exécution ».

Citons quelques EDL de TAO incomplets à notre sens.

- **ETAP-3** de l'IPPI à Moscou. L'EDL vu en démonstration est partiel, et nous n'avons pas trouvé de référence le décrivant.
- **Vermobil**. Il n'y avait pas d'EDL pour préparer les linguiciels, mais au maximum quelques script sous Linux. Par contre, il y avait un EDL pour l'intégration et la mise au point du système global, construit autour d'une structure commune (« tableau noir ») accédée par chaque module de façon distribuée.
- **Systran** (Systran). Là aussi, les développeurs sont informaticiens-linguistes. Les parties concernant les grammaires et les automates sont définies directement dans le code source. Les « composants » écrits dans les LSPL semblent n'être que les dictionnaires et les

transducteurs finis utilisés pour la morphologie. Le cycle de développement est le même pour le « cœur » de TALN que pour les interfaces et la gestion du flot de travaux, i.e. le cycle de production en génie logiciel. La partie lexicale est séparée et développée avec des outils internes : commandes, scripts pour l'indexage, filtrage d'erreurs, etc. Le code source lexical (sous plusieurs formats : texte, Excel, XML…) est compilé et encrypté par l'outil ACMulti (J. Senellart 2003). La gestion de versions est faite sous le système CVS.

2.3 Exemples de Méta-EDL : CASH et WICALE 1.0

2.3.1 CASH, pour Ariane-G5

Le système Ariane-G5 est normalement utilisable depuis une « machine virtuelle » générée par VM/ESA et accessible via http, smtp ou des sockets. Cette accessibilité ne se limite pas à l'exécution de traductions, mais s'étend à l'ensemble du développement des linguiciels, y compris à la création et à la maintenance de grammaires et de dictionnaires.

Pour faciliter cette exploitation à distance, E.Blanc a réalisé une interface hypertextuelle, CASH (Commande d'Ariane Sous Hypertexte) (E. Blanc 1996). CASH intègre plusieurs fonctions qui en font plus qu'un méta-EDL, comme l'aide à l'indexation dans les dictionnaires et l'édition graphique d'arbres et schémas d'arbres. CASH vient d'être converti de HyperCard (propre à Mac OS 9) vers une plate-forme portable (Revolution).

Dans l'exemple, on a cliqué sur la variable SUBA utilisée dans la définition de la procédure ADJ, d'où l'ouverture d'une autre fenêtre contenant la définition de cette variable.

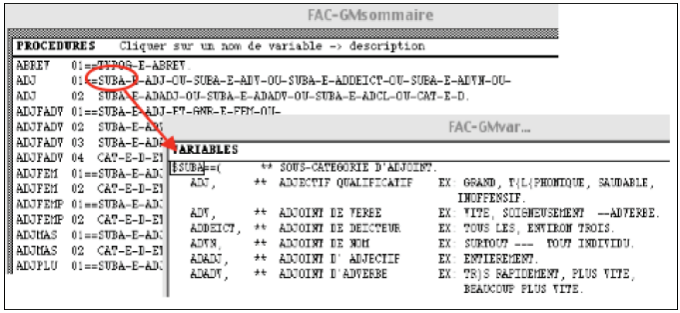


Figure 2: navigation sous CASH

2.3.2 WICALE version 1.0

Un méta-EDL minimal pour la TA, WICALE version 1.0 (V. Carpena 2004) a été construit en 2005. Il offre aux linguistes les mêmes services d'échange de données que CASH, mais aucun autre. En ce sens, il est « minimal ». Par contre, il est « générique » car il permet de travailler avec plusieurs EDL. D'un autre côté, CASH est très riche et très utile pour travailler spécifiquement avec Ariane-G5. WICALE 1.0 a été expérimenté avec les EDL d'Ariane-G5, de PILAF et d'UNL (UNL-deco du GETA).

Table 2 : comparaison entre CASH et WICALE 1.0

	Echange de commandes et données	Edition de données	Navigation dans les données	Généricité	Aide aux linguistes
CASH	oui	oui	oui	non (spécifique à Ariane-G5)	oui
WICALE 1.0	oui	non	non	oui	non

WICALE présente deux avantages principaux, la généricité et la portabilité.

- *généricité :* on peut étendre WICALE à un nouvel EDL sans écrire de code Java, mais simplement en décrivant les commandes et les données de cet EDL
- *portabilité :* elle est simplement due au fait que Java existe dans pratiquement tous les environnements logiciels actuels.

WICALE permet de définir l'architecture et les commandes d'exécution des systèmes connectés. (Des exemples pour Ariane-G5, PILAF et UNL-Deco sont donnés en annexe.)

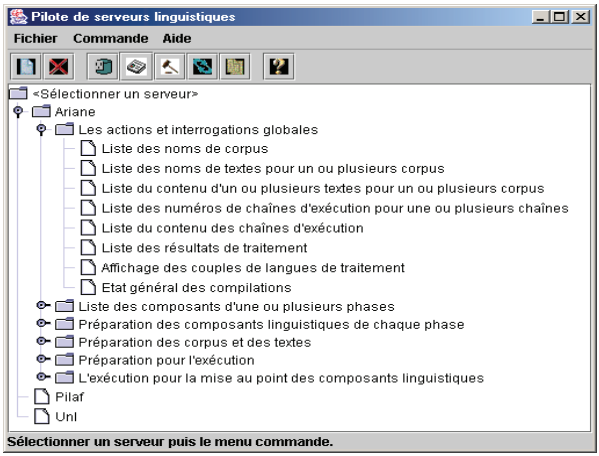


Figure 3: représentation des commandes et sous-commandes d'un EDL sous WICALE

WICALE génère l'interface correspondant aux paramètres définis dans l'architecture de chaque système. Exemple: en Ariane-G5, l'architecture est Machine>Disque>Langue>.

(Le code XML décrivant cette architecture a été supprimé de l'annexe, faute de place.)

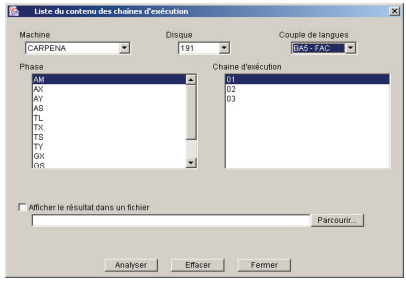


Figure 4: interface générée par WICALE 1.0

3 Enrichissement d'un méta-EDL générique: WICALE 2.0

Nous avons cherché à enrichir WICALE 1.0 en utilisant la même technique générique.

3.1 Édition en local

On utilise tout éditeur disponible (choix paramétrable), alors qu'Ariane-G5 utilise seulement XEDIT, et on s'inspire aussi d'Ariane-G5 au niveau fonctionnel : par sécurité, l'utilisateur édite toujours une copie du composant édité. Il y a deux modes, V (Visualisation) et M (Modification) : dans le premier, les modifications effectuées n'ont aucune conséquence (on avertit cependant l'utilisateur !).

Cette extension de WICALE 1.0 a été très facile à réaliser, grâce à la modularité et à la généricité du code.

3.2 Navigation

Il s'agit ici d'offrir une possibilité similaire à celle de CASH (implémentée par des scripts *ad hoc*). Nous avons proposé et implémenté une solution simple basée sur XML et inspirée de Doxygen, un outil de génération de documentation (Doxygen).

Dans cette approche, un programme analyse et marque la liaison entre les occurrences et la définition de chaque élément. De plus, il prend en compte certains commentaires spéciaux (auteur, date, résumé, …). Finalement, un générateur produit des fichiers HTML où toutes les occurrences dans le source deviennent des liens pointant vers la page contenant la définition. L'utilisateur navigue dans l'ensemble de ces fichiers HTML en utilisant n'importe quel navigateur Web.

Dans WICALE 2.0, la préparation des fichiers de navigation se passe de la même façon, en trois étapes:

- *transformation en XML du code source* des composants linguiciels, réalisée dans notre cas par le compilateur d'Ariane-Y ;
- *marquage* : parcours de la structure intermédiaire XML de chaque composant et insertion de liens entre définitions et occurrences ;
- *génération* d'un fichier html pour chaque composant, avec ajout à chaque occurrence d'un élément d'un lien vers la position de sa définition.

Voici un exemple tiré d'un composant de "définition de variables" (définition d'un jeu de décorations linguistiques) :

```
** Commentaires entre 2 étoiles et point.
** Transformation statique Jeu_1 --> Jeu_2.
-DECVAR- dv .   ** Nom du composant: DV.
-DECO- deco .   ** Nom du jeu : deco.
MT              ** Temps morphologique.
==(IMP,IPR,SPR,IPA,SPA,INF,PPR,PPA,FUT,CDL).
SEXE == (FEMININ, MASCULIN).
DGA == (SYN, ANA, NO).
…
-CVAR-
** Transformation complémentaire (procédure).
CHGMT(C;@S) ==
-SI-    MT(@S)-INC-IPR -ALORS- MT(C):=IPR;
-SNSI- MT(@S)-INC-SPR -ALORS- MT(C):=IPR;
-SNSI- MT(@S)-INC-IPF -ALORS- MT(C):=IPA;
-SNSI- MT(@S)-INC-SPF -ALORS- MT(C):=SPR;
-SNSI- MT(@S)-INC-IPA -ALORS- MT(C):=IPA;
-SNSI- MT(@S)-INC-FUT -ALORS- MT(C):=FUT;
-SNSI- MT(@S)-INC-CDL -ALORS- MT(C):=SPA;
-SNSI- MT(@S)-INC-IMP -ALORS- MT(C):=IMP;
-SNSI- SUBV(@S)-E-INF -ALORS- MT(C):=INF;
-SNSI- SUBV(@S)-E-PPR -ALORS- MT(C):=PPR;
-SNSI- SUBV(@S)-E-PPA -ALORS- MT(C):=PPA;
-FSI-.
…
-FIN-
```

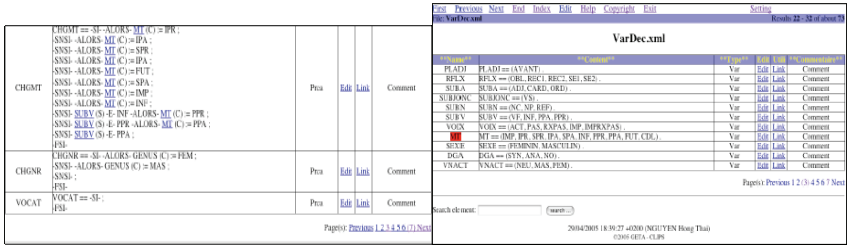


Figure 5: navigation sous WICALE 2.0

Après la génération, on a des fichiers HTML et on peut y naviguer.

4 Intégration d'une base de données lexicales multilingue

4.1 Nécessité & difficulté

4.1.1 Nécessité

Si l'on veut développer un système de TA hétérogène grâce à un EDL, il faut y centraliser de traitement des parties multilingues. En approche transfert, cela implique de traiter les grammaires et les dictionnaires de transfert, la partie lexicale étant la plus importante.

En approche par « pivot interlingue », il faut centraliser le développement des dictionnaires pivot-Li, pour chaque langue Li. G.Sérasset a développé un premier exemple d'une telle base lexicale dans le serveur UNL-deco pour traduire le site B@bel de l'UNESCO en français, espagnol, russe, et chinois (C. Boitet 2005).

4.1.2 Problèmes rencontrés lors de travaux antérieurs

- **CICC**. C'est un projet de l'ODA (Overseas Development Agency) pour la TAO « à pivot » entre japonais, chinois, thaï, malais et indonésien, financé par de grosses sociétés japonaises actives en TAO. On y tenta de développer le vocabulaire IL comme un ensemble de « concepts », identifiés uniquement à l'aide de définitions en anglais. Mais cette contrainte était trop forte (comment distinguer 2 poissons par 2 définitions ?), et il n'y avait pas de base de données partagée en accès direct.
- **UWGate**. Cette base de données lexicales centralisée pour le projet UNL donne l'accès par échange de fichiers (gzip protégé), même pour un seul article de dictionnaire. De plus, le délai d'attente est bien trop long (accusé de réception après 2 ou 3 jours ou jamais…).
- **UNL-deco**. Ce service web de déconversion d'UNL vers le français contient une base de données lexicales accessible par le Web en temps réel, mais actuellement limitée au français et à UNL. Elle n'a en fait pas été utilisée pour développer le système fra-UNL, car elle n'offre aucun outil d'aide aux travaux lexicographiques (tri, filtre, aide à l'indexage…) et elle est inextensible à N langues, au contraire de CASH+PARAX.

4.1.3 Difficulté de principe

Les problèmes décrits plus haut nous semblent provenir d'une difficulté de principe, à savoir que le problème très général de construire une base de données lexicales « universelle » pour la TAO, capable de gérer tous les aspects, de la construction des données jusqu'à l'extraction automatique de dictionnaires des modules des différents systèmes (ex. analyse morphologique de Systran, transfert lexical de Neon), est quasiment insoluble.

Non seulement les difficultés théoriques sont encore plus grandes que dans le cas d'une base lexicale multilingue « d'usage », destinée aussi à la recherche en dictionnairique multilingue, comme la base Papillon (Projet PAPILLON), mais les difficultés pratiques sont quasiment insurmontables (diversité des formats, et pire encore de la nature des informations, problèmes de droits de propriété intellectuelle (IPR)).

4.2 Une première approche, se limiter à des systèmes de TAO à pivot

L'analyse résumée ci-dessus a cependant montré qu'il devrait être possible de simplifier ce problème et d'arriver à un problème soluble en théorie, et à une réalisation utile en pratique. La simplification dont il s'agit a deux aspects :

- on considère une architecture lexicale « en étoile », ou « à pivot », ce qui revient à se limiter à des systèmes de TAO à « pivot lexical ».

- on renonce à ce que le système contrôle totalement les données lexicales, jusqu'à leur représentation « codée » dans les divers systèmes de TAO considérés.

Cette approche a déjà été réalisée et validée par la base de données lexicales PARAX (Blanc 1999). Mais cet environnement de développement, très adapté pour les manipulations lexicales, ne permet pas le travail coopératif à distance. PIVAX s'inspire donc de PARAX, mais l'étend au niveau structurel, et vise un fonctionnement distribué de type Wiki.

4.3 Vers PIVAX, une base de données lexicales contributive pour systèmes à pivot lexical

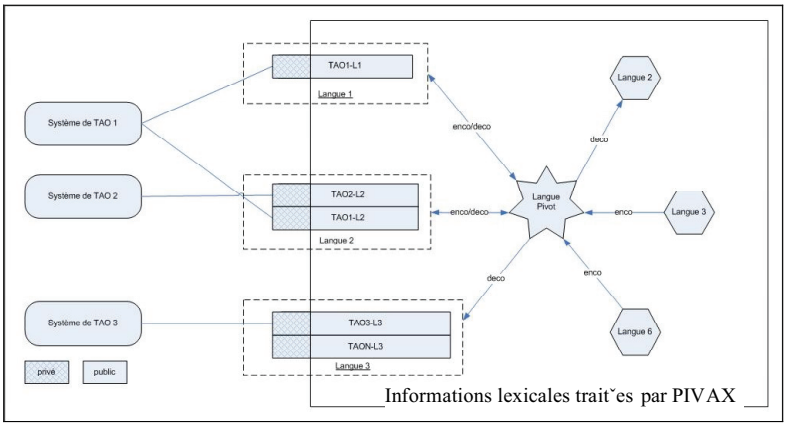


Figure 6: schéma de l'information présentée dans PIVAX

Dans PIVAX, on considère que chaque système possède ses propres linguiciels privés non gérés par PIVAX, et partage via PIVAX la partie « pivot » ainsi que sa partie « publique ». Dans la partie privée, on mettra par exemple des informations spécifiques réservées pour ce système comme les codes (morphologiques, syntaxiques et sémantiques) liés aux LSPL, les familles dérivationnelles (UL du GETA), et les formules sémantiques. La partie « pivot » contient des unités conceptuelles (IF, CATALYST) ou des acceptions interlingues (ATLAS-II, ULTRA, PIVOT, UNL), et la partie « publique » contient des lemmes ou des lexies (lemme avec indication de sens), entités par essence non propriétaires.

PIVAX sera accessible pour le travail lexicographique (humain) par une interface du type de celle de PARAX. D'autre part, PIVAX offrira une API pour la synchronisation avec divers systèmes de TAO, réalisée avec les modules existants de WICALE. Notons enfin que PIVAX est développé sur la plate-forme Jibiki de G. Sérasset, déjà utilisée pour développer la base lexicale PAPILLON, la terminologie multilingue de la Convention Alpine (Projet LexAlp), et le Grand Dictionnaire Estonien-Français GDEF (Projet GDEF).

Conclusion et perspectives

Nous avons exposé la conception d'un méta-EDL générique et d'un EDL « intégrateur », qui ne seraient pas nécessairement limités à la TAO comme WICALE. Dans un premier temps, nous avons cherché à incorporer au méta-EDL WICALE 1.0 toutes les fonctions de CASH. Concrètement, nous y avons ajouté une possibilité d'édition des composants linguiciels, puis une possibilité de navigation utilisant une compilation « légère » vers un format XML,

obtenant WICALE 1.1. D'autre part, nous avons décrit la construction en cours de PIVAX, une base de données lexicales contributive à pivot interlingue destinée à la TAO et aux autres applications multilingues (RI, traitement de contenu…). Nous espérons pouvoir présenter lors de TALN une première version de PIVAX, appliquée au développement de la base lexicale du projet UNL/U++ (français, anglais, espagnol, russe au moins).

Références

BLANC, E. (1996). Une maquette de base lexicale multilingue à pivot lexical ("acceptions multilingues"): PARAX. *Actes des quatrièmes Journées scientifiques du Réseau "Lexicologie, Terminologie, Traduction" de l'AUF, Lyon (France)*, 43-58.

BLANC, E. (1999). An interactive hypertextual environment for MT development. *MT Summit VIII, Santiago de Compostela, Galicia, Spain*, EAMT, 67-81.

BOITET, C. (1988). Dictionnaires intégrés multiusage et multicible, une première exprérience. *Colloque sur l'histoire de la terminologie, Inistitut Libre Marie Haps, Bruxelles*, 6 p.

BOITET, C. (1989). Software and lingware engineering in modern M(A)T systems. *Computational Linguistics, an International Handbook on Computer-Oriented Language Research and Applications*, Niemeyer, 670-682.

BOITET, C. and NÉDOBEJKINE, N. (1986). Towards integrated dictionary for M(a)T: motivations and linguistic organization. *Proc. COLING-86, Bonn,* 1/1, 423-428.

CARPENA, V. (2004). Interface cliente générique pour le pilotage de serveurs linguistiques. *Mémoire CNAM GETA, CLIPS, Grenoble*, 86 p.

DOXYGEN. http://www.stack.nl/~dimitri/doxygen/, accédé en 2007.

LAFOURCADE, M. (1994). Génie Logiciel pour le Génie Linguistique. *Thèse, UJF, Grenoble*.

MANGEOT-LEREBOURS, M. (1999). Accès unique à des dictionnaires hétérogènes. *VIe journées scientifiques du réseau thématique LTT de l'AUF (Lexicologie, Terminologie, Traduction)*, 311-316.

MANGEOT-LEREBOURS, M. (2001). Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue. *Thèse, UJF, Grenoble,* 279 p.

MANGEOT-LEREBOURS, M., SÉRASSET, G. (2001). Projet Papillon: architecture du serveur Web. *JST'2001 Journées Science et Technologie, National Olympic Memorial Youth Center, Tokyo, Japon* 1/1, 149-150.

NEY, H., OCH, F. J. AND VOGEL, S. (2000). Statistical Translation of Spoken Dialogues in the Vermobil System. *Proc. MSC2000*, 69-74.

NGUYEN, H.-T. (2005). Vers un "méta-EDL", puis un "EDL générique" pour la TAO. Mémoire de master recherche (M2R), UJF, Grenoble, 85 p.

NIRENBURG, S. AND FREDERKING, R. (1994). Toward multi-engine machine translation. *Proc. of the workshop on Human Language Technology, Plainsboro, New Jersey, USA*, 147-151.

PROJET PAPILLON (2003). http://www.papillon-dictionary.org/, accédé en 2007.

PROJET ARIANE-Y (2004). http://www-clips.imag.fr/geta/User/jean-philippe.guilbaud/DOCUMENTS/ARIANE-Y/ARIANE-Y-Index.html, accédé en 2007.

PROJET C-STAR (2004). http://www.c-star.org/, accédé en 2007.

PROJET UNL (1997). http://www.unl.ias.unu.edu/, accédé en 2007.

PROJET GDEF (2005). http://estfra.ee/Home.po, accédé en 2007.

PROJET LEXALP (2004). http://217.199.4.152:8080/general/lexalp/index.php, accédé en 2007.

SENELLART, J., YANG, J. AND REBOLLO, A. (2003). Technologie "Intuitive Coding" de SYSTRAN. *MT Summit IX*, 8p.

SÉRASSET, G. (1994). SUBLIM: un système universel de bases lexicales multilingues et NADIA: sa spécialisation aux bases lexicales interlingues par acceptions. *Thèse,* UJF*, Grenoble*, 194 p.

SYSTRAN (2006). https://systran.fr, accédé en 2007.

Annexe

Exemple de déclaration des commandes d'Ariane-G5, de PILAF et d'UNL-deco :

```
< !-–Description des commandes Ariane-G5 -->
<LST_SERVEUR>
 <SERVEUR>
  <nom_serveur>ARIANE-G5</nom_serveur>                        < !--Nom de serveur -->
  <classe>ServerAriane</classe>                              < !--Type de serveur -->
  <Communication>Socket</Communication>            < !--Protocole de communication-->
  <adresse>tupai.imag.fr</adresse>                        < !--Adresse du serveur -->
  <port>5768</port>                                          < !--Port du serveur -->
  <codage>iso-8859-1</codage>                          < !--Encodage des données -->
  <entete_ligne> &lt;&lt;&lt; 19283 &gt;&gt;&gt; --- premier enregistrement ---
ARIANET --- LIDIA20 ---.                                                </entete_ligne>
  <fin_ligne> &lt;&lt;&lt; 19283 &gt;&gt;&gt; --- premier enregistrement --- ARIANET
--- LIDIA20 ---.                                                          </fin_ligne>
 </SERVEUR>
```

```
 <SERVEUR>
  <nom_serveur>PILAF</nom_serveur>
  <classe>ServerHttp</classe>
    <communication>Http</communication>
    <adresse>http://clips.imag.fr/cgi-
bin/pilaf/</adresse>
    <port></port>
    <codage>iso-8859-1</codage>
    <entete_ligne></entete_ligne>
    <fin_ligne></fin_ligne>
 </SERVEUR>
```

```
<SERVEUR>
  <nom_serveur>UNL</nom_serveur>
  <classe>ServerUnl</classe>
    <communication>Socket</communication>
    <adresse>tupai.imag.fr</adresse>
    <port>5768</port>
    <codage>iso-8859-1</codage>
    <entete_ligne></entete_ligne>
    <fin_ligne></fin_ligne>
</SERVEUR>
</LST_SERVEUR>
```

Description de la commande d'Ariane-G5 qui demande la liste des noms des corpus. En natif, sa forme est **LISNOMCORP** (Terminal | Imprimante | TI).

```
<COMMANDE num_cde="1">
  <num_cde>1</num_cde>
  <nom_cde>LISNOMCORP</nom_cde>
  <intitule_cde>Liste des noms de
corpus</intitule_cde>
  <PARAMETRE_SAISIE> </PARAMETRE_SAISIE>
  <SYNTAXE>
    <ETAPE>
    <mot_cle>TRAIT = LISNOMCORP (*)
</mot_cle>
    <num_param></num_param>
    <expression></expression>

  <separateur>&retour_chariot;</separateur
> ...

  <saisie_obligatoire>false</saisie_obliga
toire>
    </ETAPE>
  </SYNTAXE>
```

```
 <RESULTAT>
  <resultat_OK>-> Tout est
O.K.</resultat_OK>
  <resultat_type>-> Tout est
O.K.</resultat_type>
    <ETAPE>
    <nom_methode>find</nom_methode>
    <expr_deb>Liste des noms de
corpus</expr_deb>
    <expr_corps> [A-Za-z0-9]*
</expr_corps>
    <expr_fin>-LISTE TERMINEE-
</expr_fin>

  <expr_concat>&retour_chariot;</expr_c
oncat>

  <expr_remplacement></expr_remplacemen
t>
    </ETAPE>
  </RESULTAT>
</COMMANDE>
```

Aides à la navigation dans un corpus de transcriptions d'oral

Frederik CAILLIAU[1, 2], Claude DE LOUPY[3]
[1] LIPN – Institut Galilée – Université Paris-Nord,
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
[2] Sinequa Labs – 51 rue Ledru-Rollin, 94200 Ivry-sur-Seine
[3] Syllabs – 3 rue Castex, c/o Agoranov, 75004 Paris
cailliau@sinequa.com, loupy@syllabs.com

Résumé. Dans cet article, nous évaluons les performances de fonctionnalités d'aide à la navigation dans un contexte de recherche dans un corpus audio. Nous montrons que les particularités de la transcription et, en particulier les erreurs, conduisent à une dégradation parfois importante des performances des outils d'analyse. Si la navigation par concepts reste dans des niveaux d'erreur acceptables, la reconnaissance des entités nommées, utilisée pour l'aide à la lecture, voit ses performances fortement baisser. Notre remise en doute de la portabilité de ces fonctions à un corpus oral est néanmoins atténuée par la nature même du corpus qui incite à considérer que toute méthodes permettant de réduire le temps d'accès à l'information est pertinente, même si les outils utilisés sont imparfaits.

Abstract. In this paper we evaluate the performances of navigation facilities within the context of information retrieval performed on an audio corpus. We show that the issues about transcription, especially the errors, lead to a sometimes important deterioration of the performances of the analysing tools. While the navigation by concepts remains within an acceptable error rate, the recognition of named entities used in fast reading undergo a performance drop. Our caution to the portability of these functions to a speech corpus is attenuated by the nature of the corpus: access time to a speech corpus can be very long, and therefore all methods that reduce access time are good to take.

Mots-clés : évaluation, moteur de recherche, corpus oral.

Keywords: evaluation, search engine, speech corpus.

1 Introduction

Les corpus oraux font de plus en plus partie de notre quotidien, aussi bien à travers le web que dans notre environnement professionnel. Leur taille est dans une phase de très forte croissance du fait de la généralisation des podcasts. Face à la masse grandissante de données disponibles et les grands progrès constatés dans les technologies de transcription depuis les 15 dernières années, la recherche à l'intérieur de ces enregistrements s'impose. En particulier, les techniques d'aide à la navigation appliquées aux corpus écrits devraient être particulièrement utiles du fait du temps nécessaire à l'écoute d'une émission entière. Il a été montré que les performances des outils de transcriptions ont une influence relativement faible sur les

performances des moteurs de recherche sur l'audio [Allen, 2002]. Mais ces performaces de transcriptions ont-elles un impact important sur les aides à la navigation ?

Cette évaluation s'inscrit dans une série de travaux menés depuis les années 90 comme la BNN (Merlino *et al.*, 1997), Speechbot (Van Thong *et al.*, 2002), SCAN (Choi *et al.*, 1999). Des outils d'aide à la navigation pour l'audio ont déjà été testés (Anick & Tipirneni, 1999) mais concernent des fonctionnalités moins évoluées que ce qui est actuellement utilisé pour l'écrit.

Le présent article se place dans le cadre et à la suite du projet AudioSurf[1] dont le but était de créer une plate-forme d'indexation de l'audio et, en particulier, un moteur de recherche sur l'audio ayant les mêmes fonctionnalités qu'un moteur de recherche sur le texte. Les moteurs de recherche sur les textes écrits ont fait de grands progrès depuis quelques années en incluant des fonctionnalités d'aide à la navigation qui permettent de donner des informations complémentaires à l'utilisateur, de lui permettre de spécifier sa requête et d'interagir avec le système.

Nous présentons ici une application du moteur Intuition de Sinequa à l'indexation de corpus oraux transcrits à l'aide de l'outil du LIMSI et de Vecsys (Gauvain *et al.*, 2000) et les conséquences des particularités de tels corpus sur les fonctionnalités d'aide à la navigation. En section 2, nous décrivons le moteur de recherche Intuition, le principe des aides à la navigation ainsi que l'évaluation de leur apport. La section 3 décrit le corpus de transcriptions, ses particularités ainsi que les implications de ces particularités sur les performances de l'outil. Enfin, en section 4, nous présentons les résultats des évaluations que nous avons menées.

2 La plateforme Intuition

2.1 Présentation

Intuition est une plateforme de recherche d'information développée par Sinequa[2], constituée d'un moteur de recherche et d'interfaces de navigation. Elle repose sur des traitements linguistiques, statistiques et sémantiques qui augmentent la pertinence des documents trouvés et accélèrent la recherche des utilisateurs (cf. section 2.2).

La figure 1 présente l'interface du moteur de recherche telle qu'elle a été conçue pour le corpus audio. Globalement, cette interface est similaire à celle sur les textes écrits. Certains éléments ont cependant été ajoutés comme l'accès direct à l'écoute du passage, sa durée, etc.

---

[1] Le projet AudioSurf a été financé dans le cadre du Réseau National des Technologies Logicielles (appel RNTL 2002). Il avait comme partenaire Sinequa[1] (leader), la société Vecsys[1], le LIMSI[1] et le partenaire valideur Radio France.

[2] Pour plus d'informations : http://www.sinequa.com/

Sur le volet de gauche, apparaissent des listes de concepts[3], d'entités (noms de lieux, d'organisations et de personnes). Ces éléments sont contextuels par rapport à la requête (Crestan & Loupy, 2004) et permettent à l'utilisateur de la préciser. Un simple clic sur une des entités permet de relancer une requête demandant des documents répondant à la requête précédente et contenant le terme choisi. L'extraction des entités, en combinaison avec un équilibrage statistique, se transforme alors en générateur de filtres à la volée qui permet de restreindre rapidement le nombre de documents de la liste des réponses.

L'extraction des entités est faite à partir de grammaires locales écrites sous forme de transducteurs, qui prennent en compte les résultats d'un étiquetage morphosyntaxique et d'une lemmatisation. Plus d'informations sur les ressources linguistiques utilisées dans ces traitements peuvent être retrouvées dans Cailliau (2006).

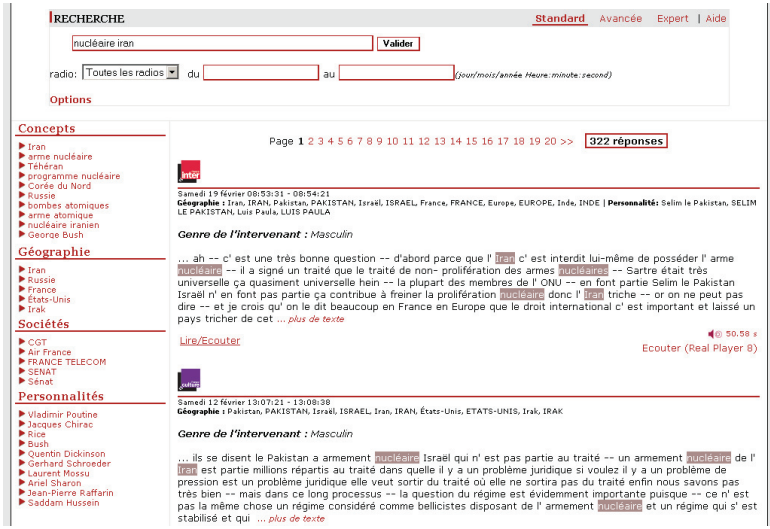


Figure 1 : Interface du moteur de recherche sur le corpus audio

L'aide à la lecture est une deuxième application des entités visible pour l'utilisateur. Elle consiste à mettre en couleur les différentes entités nommées qui ont été identifiées à l'intérieur d'un document afin de favoriser une lecture rapide par le repérage des passages importants. Par exemple, les personnes seront visualisées en rouge, les lieux en bleu, etc. Il est ainsi possible de repérer très vite ce dont parle le document. Pour l'évaluation, nous nous concentrerons sur le rappel et la précision en reconnaissance des personnes par le système.

Les interfaces décrites se complètent par un autre type de navigation, qui ne sera pas évalué dans cet article : la fonction des documents similaires. Elle permet de retrouver des documents sémantiquement proches de celui que l'utilisateur vient de regarder.

[3] Appelés parfois aussi *termes associés*.

Nous évaluerons l'impact, sur ces fonctionnalités d'aide à la navigation, du passage à des corpus oraux dans la section 4.

2.2 Évaluation du principe de navigation

Le principe de navigation utilisé ici a été validé sur l'écrit [Crestan & Loupy, 2004]. Nous avons effectué une analyse mettant en jeu :

- 775 000 articles issus du journal *Le Monde* (années 1989 à 2002) ;
- 6 interfaces différentes utilisant l'une ou l'autre des fonctions de navigation ;
- 18 requêtes dont 12 de type recherche documentaire (traductions de requêtes provenant de TREC-6, ad'hoc [Voorhees & Harman, 1997]) et 6 requêtes factuelles (traductions de requêtes provenant de TREC-11, question/answering [Voorhees, 2003]) ;
- 6 personnes de formation et intérêts différents ayant pour instruction de passer exactement 10 mn par requête pour retrouver le maximum de documents pertinents. Chaque document visualisé devait être classé pertinent ou non pertinent par l'utilisateur.

Les résultats ont été très satisfaisants puisque l'interface donnant accès à toutes les aides à la navigation a permis :

- de diminuer le temps d'accès au premier document pertinent par deux en moyenne (248 s → 122 s) ;
- d'augmenter presque par deux en moyenne le nombre de documents pertinents retrouvés (3,83 → 6,56) ;
- de diminuer très significativement le nombre de documents non pertinents visualisés (7,17 → 4,28).

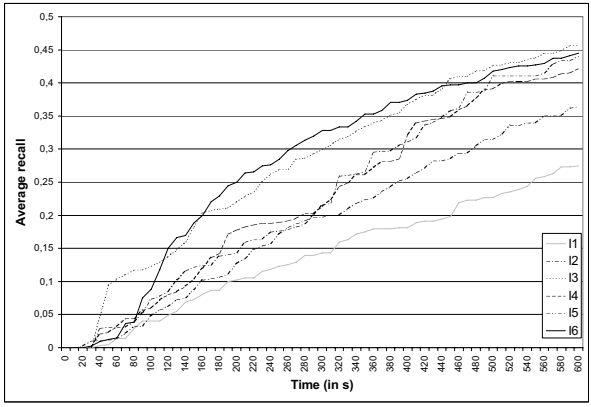


Figure 1 : Évaluation de l'apport des aides à la navigation en prenant compte du temps

La courbe précédente montre la progression du nombre de documents pertinents récupérés en utilisant les différentes interfaces. La courbe I6 (interface utilisant toutes les fonctionnalités d'aide à la navigation) obtient des résultats très au-dessus de la courbe I1 (interface basique).

Ces expériences ont donc montré l'intérêt de ces aides à la navigation sur un corpus écrit. Un corpus oral présente en revanche des difficultés pouvant réduire l'utilité de telles fonctionnalités.

3 Difficultés apportées par la transcription

Le corpus est constitué de 1048 fichiers au format xml, représentant chacun une heure de transcription automatique. Ils couvrent l'ensemble des émissions radio de France Culture et de France Inter dans la période du 6/2/05 au 28/2/05. L'unité habituelle d'indexation dans un contexte de corpus écrit est le fichier, qui correspond dans la majorité des cas aussi à une unité thématique. La notion de document a donc dû être redéfinie puisque plusieurs émissions sont présentes dans une heure de radio.

La transcription issue de l'outil de Vecsys et du LIMSI (Gauvain *et al.*, 2000) n'est pas un texte conforme à ceux habituellement traités à Sinequa. La Figure 2 montre un passage transcrit par cet outil.

Intervenante féminine

au troisième jour de sa **visite** officielle en **Chine** le président français Jacques **Chirac** a une nouvelle fois salué les relations de confiance qui unissent les deux pays -- lors d' une conférence de presse de clôture le chef de l' état a fait le bilan de son déplacement -- il a souligné la proximité de vues entre Paris et Pékin sur les grands dossiers internationaux du moment -- depuis la **Chine** le président est aussi intervenu dans le débat qui a cours actuellement dans l' hexagone concernant la retransmission télévisée des conseils des ministres une proposition formulée par le chef du gouvernement Dominique de Villepin et à laquelle le président s' est dite opposée au niveau économique et la **visite** de Jacques **Chirac** en **Chine** a permis à Airbus et à Alstom de rafler plusieurs gros contrats le chef de l' état a également posé ce vendredi -- la première pierre d' une usine PSA Peugeot Citroën avoue Anne la grande ville industrielle du centre du pays -- depuis peu Anne justement Jacques **Chirac** a évoqué le dossier du nucléaire iranien -- le chef de l' état s' est prononcé en faveur de sanctions adaptées et provisoire et irréversible contre Téhéran -- en cas d' échec du dialogue avec les Européens sur l' enrichissement de l' uranium

Intervenant masculin

il est évident -- et l' ensemble de -- la communauté internationale reconnaît notamment la Chine mais aussi la Russie -- les Européens les États-Unis -- que l' ambition clairement affichée -- par l' Iran n' est pas compatible avec une certaine idée que nous nous faisons de la non- prolifération -- en raison des dangers qu' elle représente sa propre représentera prolifération

Figure 2 : Exemple de transcription (les mots en gras sont ceux présents dans la requête d'origine)

Nous voyons ici, des caractéristiques classiques d'une transcription automatique :

– Chaque fichier xml est structuré par les tours de parole. Ceux-ci ne représentent pas forcément une unité thématique, mais ils ont été indexés comme des documents faute d'un meilleur découpage. Les balises des tours de parole comportent des attributs avec l'heure de début et de fin du tour de parole. Ces informations sont exploitées dans la maquette pour retrouver la partie du fichier audio qui y correspond. D'autres attributs non exploités sont l'identifiant de la personne qui parle et son sexe. Un flux de parole n'est pas aussi propre qu'un article de journal. Les locuteurs peuvent se couper la parole, parler en même temps ce qui rend la transcription très aléatoire. La qualité de transcription peut être très différente d'un locuteur à un autre selon la façon d'articuler, l'accent ou le fait que le journaliste peut être sur son plateau de radio alors que l'interviewé est au téléphone (d'où une qualité de son très mauvaise).

– Le texte ne comporte aucune ponctuation. Les seules ponctuations présentes sont les deux traits qui indiquent une pause, une respiration, mais auxquels on ne peut attribuer de sémantique ou de syntaxe significative pour nos traitements. Les majuscules de début de phrase ne sont pas non plus données. L'unité phrastique est donc complètement absente et cède sa place au tour de parole. Nous avons tenté d'effectuer des adaptations de nos modèles statistiques pour prendre en compte ce phénomène, mais il aurait fallu un étiquetage de corpus pour pouvoir l'effectuer de manière correcte. Du fait de l'absence d'un tel corpus étiqueté, les expériences, utilisant des corpus normaux transformés pour ressembler à du corpus oral n'ont pas été probantes. A cause de l'absence de ponctuation et du fait de la syntaxe propre à l'oral, les transcriptions, même si elles sont de très bonne qualité, sont souvent difficilement lisibles. L'écoute des morceaux sélectionnés s'impose pour une bonne compréhension. Un tour de parole dans un journal se termine par exemple souvent par le nom de la personne qui va prendre la parole juste après dans la suite du bulletin sans aucune transition : « […] *une gauche à réunifier dans un bel ensemble Frédéric Pommier* ».

– La transcription comporte l'ensemble des disfluences, hésitations, répétitions, faux départs, etc. propres à l'oral et présente donc souvent des différences importantes par rapport à un texte écrit.

– Il y a des erreurs de transcription. La transcription de la Figure 2 est excellente mais comporte malgré tout quelques erreurs. Ainsi, à la 9ème ligne, le logiciel de transcription a écrit « *avoue Anne* » au lieu de « *à Wuhan* » (ville n'étant pas présente dans le lexique du système). Les transcriptions sont en général de très bonne qualité : le Word Error Rate (WER) sur les émissions radio a été évalué en 2001 à 20% sur le type de corpus qui nous intéresse ici (Gauvain *et al.*, 2001) mais les modèles ont été améliorés depuis et ont été évalués à 11,9% de WER pendnat la campagne ESTER (Galliano et al., 2005). D'après nos observations, les erreurs relevées dans le corpus donné sont dues à la présence d'un bruit ou d'une musique de fond, à des lacunes lexicales ou au non-branchement de la détection de la langue. Pour ce dernier cas, il arrive qu'une personne parle en anglais et qu'un interprète effectue alors la traduction. Un grand nombre d'erreurs est alors généré. Dans un esprit un peu différent, les chiffres peuvent être transcrits en lettres, ce qui est le cas pour certaines années ou dans l'exemple suivant : « […] *une petite baisse de zéro zéro neuf pour-cent à quatre mille cinq points* […] ». Il est bien sûr possible de traiter facilement ce dernier point mais des particularités de ce type impliquent des ajouts de modules.

L'ensemble de ces points conduit à un certain nombre d'erreurs et de problèmes pour les fonctionnalités qui suivent, en particulier les traitements linguistiques.

4 Évaluation

Nous avons mis en place la plate-forme Intuition avec un corpus oral sans aucune adaptation des traitements décrits dans 2. Nous mesurerons leurs performance et robustesse sur un corpus oral à travers deux fonctions principales d'Intuition : la navigation par les concepts et les entités nommées d'une part et l'extraction des entités dans les documents qui servent à l'aide à la lecture d'autre part. Ce qui est évalué ici n'est pas le WER de la transcription mais son impact sur l'aide à la navigation.

4.1 Évaluation de la navigation

4.1.1 Navigation par concepts

A partir d'un jeu de requêtes existant issues de logs d'un client de Sinequa, 40 requêtes ont été sélectionnées auxquelles au moins 50 documents dans le corpus répondent. Ces requêtes, de un à quatre mots, n'ont subies aucune modification (casse, orthographe, etc.). Elles posent des questions sur des noms de personnes (*saddam* ; *mahmoud abbas* ; …), des questions sur des noms de personnes en association avec un concept (*sistani irak* ; *sharon rice paix* ; …), des questions thématiques (*fatah* ; *armes nucléaires* ; …) ou factuelles pour obtenir une information précise (*chiites élections irak* ; *tgv Paris strasbourg* ; …). Le jeu complet est présenté en annexe.

Afin de pouvoir comparer les résultats de l'évaluation sur le corpus oral aux performances posées sur l'écrit, nous avons fait les mêmes tests sur un corpus écrit de type presse, composé de 21984 fichiers xml pour une totalité de 81,1 Mo. Le corpus oral est donc celui présenté en section 3.

Nous avons évalué les concepts qui sont extraits en fonction d'une requête. Cette évaluation est basée sur leur structure, c'est-à-dire que nous avons cherché à savoir s'ils étaient bien formés. Le but de cet article étant d'analyser l'impact des particularités de la transcription, la pertinence des concepts par rapport à la requête n'est pas évaluée.

L'évaluation elle-même porte sur les 40 premiers concepts rapportés par le système. Elle a été effectuée par 3 personnes ayant une compétence en linguistique. Pour chaque concept présenté, il était demandé de noter s'il était ou non bien formé. La figure suivante montre l'évolution des erreurs dans les concepts extraits.

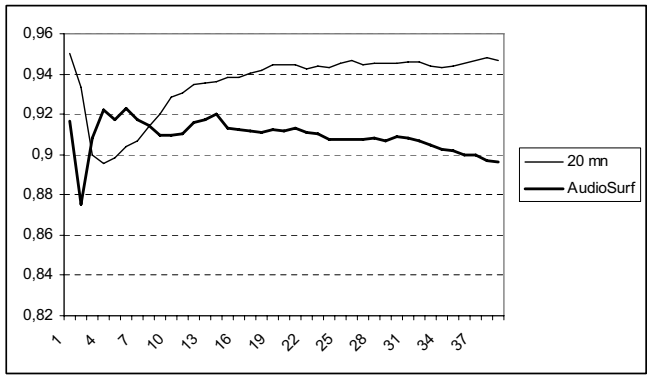


Figure 3 : Évaluation du nombre de concepts bien formés extraits d'un corpus audio (AudioSurf) et d'un corpus écrit (20 mn)

On peut voir que les concepts mal formés sont plus nombreux sur les transcriptions que sur des textes normaux comme nous pouvions nous y attendre. Le taux d'erreur moyen sur le texte est de 5% et de 10% sur les transcriptions.

Le phénomène de forte décroissance en début de courbe est dû à de mauvaises analyses des concepts et non à une mauvaise transcription (on peut voir qu'elle apparaît aussi sur le texte). Il s'agit de noms de lieux ou de personnes du Moyen Orient pour lesquels les transducteurs ne sont pas assez robustes. 16 requêtes sur 40 portent sur cette région du monde et des concepts comme *Char el* (au lieu de *Charm el Cheikh*) sont fréquemment extraits et se trouvent en tête de liste car le concept complet est très pertinent.

La croissance du nombre d'erreurs est assez forte (5 points) lorsque l'on passe aux transcriptions. Cette décroissance provient d'erreurs comme « *sa patte héros* » au lieu de *Zapatero*, « *Traque Tommy* » pour *trachéotomie* ou « *Langeais Luce* » à la place de « *l'Angélus* ». Néanmoins, nous restons dans un ordre d'erreurs acceptable (un concept sur 10 mal formé), c'est-à-dire qu'il est visible et bien présent mais non préjudiciable à la navigation.

4.1.2 Navigation par entités nommées

L'évaluation de la navigation par entités nommées n'était pas pertinente dans le contexte présent. Les lieux et les entreprises sont extraits par listes (avec quelques contextes restrictifs). Sur les entreprises, seules des choses correctes sont renvoyées et les manques viennent plutôt de l'incomplétude des listes utilisées. Pour la géographie, le rappel et la précision sont très bons mais certaines erreurs récurrentes apparaissent avec « France deux » et « France inter » dont le premier terme est reconnu comme le pays.

4.2 Évaluation de l'aide à la lecture

Cette évaluation est semblable à l'étude qu'ont faite Kubala *et al.* (1998). Les émissions de France Culture comportant peu d'entités nommées, nous avons choisi comme échantillon deux heures d'émission de France Inter. L'étude a porté sur la première demi-heure de chacune de ces émissions, car c'est la partie qui est la plus dense en entités nommées.

L'identification des entités nommées est fortement liée à la présence de ces entités dans les lexiques utilisés pour la reconnaissance de la parole. Si le nom propre est inconnu de ces lexiques, les mots en question sont remplacés par des mots communs, ce qui rend impossible toute détection par les grammaires d'extraction.

Le tableau suivant présente une évaluation du rappel et de la précision de la reconnaissance des personnes dans 3 contextes :

- La transcription automatique c'est-à-dire sans se préoccuper des mauvaises transcriptions. Ainsi, si une personne est citée à l'oral mais que la transcription en la fait pas apparaître (elle se trompe), elle n'est pas prise en compte dans le calcul.
- La confrontation à l'oral : si une entité est mal transcrite, elle sera comptabilisée quand même, ce qui fait chuter le rappel. Nous avons donc corrigé la transcription automatique pour effectuer cette évaluation.
- La transcription manuelle : afin d'évaluer l'impact des erreurs de transcription, nous avons corrigé manuellement celle-ci et repassé l'extraction des entités afin de réévaluer la précision et le rappel sur une transcription jugée sans erreur par le transcripteur humain.

	transcription automatique	confrontation à l'oral	transcription manuelle
Précision	0,90	0,90	0,91
Rappel	0,73	0,65	0,80

Les erreurs de transcription mises en cause concernent des entités comme Mahmoud Abbas qui sont reconnues sous un autre nom (« *le dirigeant palestinien Marc Mbouda basses annoncent* »), des configurations différentes d'écriture comme pour « Jean Paul deux » où ce cas d'écriture n'a pas été prévu dans les transducteurs, etc.

On constate que les erreurs de transcription ont pour conséquence une chute du rappel de 15 points par rapport à une transcription manuelle. Ce chiffre est très important et nous conduit à penser que l'utilisation de cette fonctionnalité sur un corpus oral n'est peut-être pas pertinente. Néanmoins, les transcriptions ne sont pas faites pour être lues mais plutôt pour déterminer de quoi parle un texte et si l'on veut aller plus loin en écoutant l'émission ou non. Tous les éléments permettant d'aider l'utilisateur à appréhender plus vite l'intérêt d'une émission sont intéressants dans ce contexte. Il faudrait une évaluation de navigation avec utilisateur comme celle présentée en section 2.2 pour pouvoir conclure.

5 Conclusion et perspectives

Le but de notre étude était d'évaluer la portabilité sur du corpus oral des traitements faits habituellement sur l'écrit. En ce qui concerne la navigation par concepts, nous avons constaté une dégradation significative mais tout à fait acceptable au perçu des utilisateurs. Les performances de l'extraction des entités nommées sur les documents du corpus oral sont bien faibles en rappel, mais la précision et le rappel sont en même temps déjà un apport pour la fonctionnalité visée. Dans les systèmes qui traitent de l'oral, toute amélioration qui réduit le temps d'accès à un morceau précis est un gain pour l'utilisateur. D'autres expériences qui mettent l'utilisateur au centre de l'évaluation sont à mener, justement pour mesurer si l'apport en efficacité est comparable à celui constaté sur le texte.

Remerciements

Les auteurs tiennent à remercier Mélodie Soufflard pour son travail d'analyse et d'étiquetage.

Références

ALLEN J. (2002). Perspectives on Information Retrieval and Speech. In Information Retrieval Techniques for Speech Applications, Coden, Brown, and Srinivasan (Eds.)..

ANICK, P.G., TIPIRNENI, S. (1999). The paraphrase search assistant: terminological feedback for iterative information seeking. Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval. SIGIR '99. ACM Press, New York, NY, pp. 153-159.

CAILLIAU F. (2006). Un modèle pour unifier la gestion de ressources linguistiques en contexte multilingue. Actes de TALN 2006.

Choi J., Hindle D., Pereira F., Singhal A., Whittaker S. (1999). Spoken content-based audio navigation (SCAN). Proceedings of the ICPhS-99.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F., GRAVIER G. (2005). The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. Proceedings of the European Conf. on Speech Communication and Technology.

GAUVAIN J.L., LAMEL L., ADDA G., ADDA-DECKER M., BARRAS C., CHEN L., KERCADIO Y. DE (2001). Processing Broadcast Audio for Information Access'. ACL 39th annual meeting, pp. 2-9.

GAUVAIN J.L., LORI L., ADDA G. (2000). Transcribing broadcast news for audio and video indexing. Communications of the ACM, vol. 43, n° 2, pp. 64-70.

KUBALA F., SCHWARTZ R., STONE R., WEISCHEDEL R. (1998). Named entity extraction from speech. Proceedings of DARPA Broadcast News Transcription and Understanding. Workshop, Lansdowne, VA.

LOUPY C. DE, CRESTAN E. (2004). Browsing Help for Faster Document Retrieval. Actes de *Coling*.

MERLINO, A., MOREY, D., MAYBURY, M. (1997). Broadcast news navigation using story segmentation. *Proceedings of the Fifth ACM international Conference on Multimedia*, MULTIMEDIA '97. ACM Press, New York, NY, pp. 381-391.

VAN THONG J.M., MORENO P.J., LOGAN B., FIDLER B., MAFFEY K., MOORES, M. (2002). SpeechBot: An Experimental Speech-based Search Engine for Multimedia Content on the Web. IEEE Transactions on Multimedia, Vol 4, Nr. 1.

Annexe : liste des requêtes

1	nucléaire iran	15	Paris	29	tgv Paris strasbourg
2	chiites irak	16	Californie	30	chiites élections irak
3	russie gaz	17	Irlande	31	mur palestine
4	explosions de gaz Paris	18	Irak	32	cessez le feu intifada
5	attentat madrid	19	ONU	33	vote constitution européenne
6	mahmoud abbas	20	OTAN	34	fatah
7	aubenas	21	chirac en chine	35	forum mondial
8	Saddam	22	bush syrie	36	kyoto
9	Hussein	23	Poutine Rice	37	djihad
10	Chirac	24	assassinat hariri	38	armes nucléaires
11	Bush	25	moubarak et abdallah	39	chomage
12	jean paul deux	26	sida new york	40	grippe pape
13	Éyadéma	27	sistani irak		
14	Jean-Pierre Raffarin	28	sharon rice paix		

Session

Syntaxe

Une grammaire du français pour une théorie descriptive et formelle de la langue

Marie-Laure GUÉNOT*
Signes, Université Michel de Montaigne Bordeaux 3
marie-laure.guenot@u-bordeaux3.fr

Résumé. Dans cet article, nous présentons une grammaire du français qui fait l'objet d'un modèle basé sur des descriptions linguistiques de corpus (provenant notamment des travaux de l'*Approche Pronominale*) et représentée selon le formalisme des *Grammaires de Propriétés*. Elle constitue une proposition nouvelle parmi les grammaires formelles du français, participant à la mise en convergence de la variété des travaux de description linguistique, et de la diversité des possibilités de représentation formelle. Cette grammaire est mise à disposition publique sur le *Centre de Ressources pour la Description de l'Oral* en tant que ressource pour la représentation et l'analyse.

Abstract. In this paper I present a grammar for French, which is the implementation of a linguistic model based on corpus descriptions (notably coming from *Approche Pronominale*) and represented into the *Property Grammars* formalism. It accounts for a new proposition among formal grammars, taking part into the works that aim to promote convergence between the various researchs of descriptive linguistics and the diversity of formal representation possibilities. It is freely available on the *Spoken Data Resource Center* (CRDO), as a representation and analysis resource.

Mots-clés : développement de grammaire, ressource pour le TAL, grammaire du français, syntaxe, linguistique formelle, linguistique descriptive, grammaires de propriétés (GP).

Keywords: grammar development, resource for NLP, French grammar, syntax, formal linguistics, descriptive linguistics, property grammars (PG).

1 Introduction

Nous présentons ici une nouvelle ressource pour le TAL et pour la linguistique descriptive et formelle : une grammaire du français à large couverture s'intéressant à l'écrit et à l'oral, basée sur des positions théoriques originales et formalisée en *Grammaires de Propriétés* (Blache, 2005) (ci-après GP). Cette grammaire a été conçue dans le but de faire davantage converger la variété des descriptions linguistiques et la richesse des possibilités formelles offertes, afin de proposer à la fois d'élargir les points de vue théoriques des grammaires formelles, et de valider les descriptions linguistiques utilisées.

*Je tiens à remercier très chaleureusement Philippe Blache, sans qui cette ressource n'aurait pu voir le jour.

Une théorie descriptive et formelle. Le développement de cette ressource est basé sur la mise au point d'une *théorie de la langue*, que l'on a orientée à la fois vers la linguistique descriptive et vers la linguistique formelle.

La linguistique descriptive construit des hypothèses explicatives à partir de l'observation des productions de langue ; celles-ci constituent la matière première de la grammaire. Il s'agit là, en premier lieu, d'opérer un choix parmi les théories, les méthodes, et les descriptions linguistiques que l'on souhaite retenir, puisque l'on sait que celles-ci présentent une variété considérable, aussi bien concernant les phénomènes auxquelles la discipline s'intéresse, la couverture des travaux, et la diversité des approches théoriques sous-jacentes. L'une des conséquences évidentes de cette variété est que toutes les descriptions ne sont pas compatibles entre elles ; en outre, puisque les descriptions ne sont pas systématiquement envisagées sous l'angle de leur cohésion générale, de leur interdépendance, il arrive que certaines d'entre elles répondant pourtant à une même théorie et une même méthodologie, engendrent quand on entreprend de les rassembler en une grammaire une certaine hétérogénéité, voire parfois des contradictions. Partant de cet état de fait, nous avons donc voulu dans notre grammaire *expliciter* les mécanismes mis en œuvre par les descriptions utilisées, et en proposer un ensemble *homogène* et *cohérent*.

La linguistique formelle, quant à celle, propose des modes d'expression logico-mathématique pour représenter les régularités linguistiques et leur fonctionnement, à partir desquels on peut élaborer des *modèles* de grammaire, *i.e.* des squelettes d'organisation des informations et de leur analyse. Or la discipline connaît elle aussi une importante variété des propositions, qui s'explique par le fait que les motivations sous-jacentes aux modèles peuvent être très différentes, par les domaines de recherche qui peuvent en être à l'origine, et par l'ensemble de postulats propre à chaque modèle et en faisant l'originalité. Tout ceci offre au linguiste une remarquable gamme de possibilités de représentation. Toutefois, toutes les pistes envisagées dans le cadre de la linguistique (strictement) descriptive ne sont pas, et dans certains cas même ne *peuvent* pas être représentées formellement, en dépit de la richesse des propositions. Nous avons donc souhaité également pour notre grammaire, proposer un modèle qui *prenne en compte* toute une partie de la linguistique descriptive qui n'a pas encore été formalisée, pour en retour *valider* la pertinence du formalisme linguistique utilisé en montrant qu'il permet de représenter de telles informations.

La *ressource* que nous présentons ici est basée sur un *modèle*, lui-même élaboré à partir d'un ensemble de *positions théoriques* (§2.1) et du choix d'un *formalisme* (§2.2) ; après avoir introduit ceci, nous présenterons les points les plus originaux de la grammaire développée pour le français (§3). Nous aborderons (§4) sa mise à disposition publique et évoquerons d'autres ressources complémentaires à diffuser dans le même cadre. Enfin, nous conclurons sur les futurs développements de la grammaire.

2 Le modèle

2.1 Aspects théoriques

Une approche non générative. Là où les approches génératives (*Generative-Enumerative Syntax*), les plus répandues actuellement, définissent une langue comme une *liste* récursivement énumérable d'éléments obtenus à partir d'un nombre fini de règles, et définissent une grammaire comme un *outil* permettant de générer toutes et uniquement les phrases d'une langue, les

théories basées sur des modèles (*Model-Theoretic Syntax*) considèrent la grammaire comme un ensemble de *contraintes* portant sur la structure des expressions de langue, et la langue comme un ensemble par définition infini, tenant ainsi compte de la variabilité des degrés de grammaticalité et d'interprétabilité, ainsi que de la créativité linguistique (Pullum & Scholz, 2001). Pour ces raisons, le modèle de que l'on propose fait partie de cette seconde classe.

Une grammaire non présomptive. Nous adoptons une position qui consiste à ne pas anticiper sur la *structure externe* d'une construction au sein de la description de sa *structure interne* (Deulofeu, 2006) : autrement dit, on ne présume pas des relations qu'une construction entretient avec les autres éléments de l'énoncé, uniquement sur la base des relations qu'entretiennent ses constituants. La conséquence de ceci qui tranche le plus avec les autres grammaires formelles actuelles, est que la nôtre n'est *pas lexicalisée* : en effet, s'il est indéniable que le lexique contraint effectivement en partie la structuration syntaxique, comme le font par exemple les restrictions de sélection lexicale ou la valence des objets, nous refusons pour autant de radicaliser la vision lexicaliste et de systématiser l'expression des informations syntaxiques dans le lexique.

Un traitement non modulaire. A l'observation des corpus, on constate que l'information y est éparse et inconstante (Blache, 2004) : elle est dispersée à travers les domaines (phonétique, morphologie, syntaxe, sémantique, *etc.*), et pour chaque domaine elle est présente en quantité (de même qu'en qualité) variable. Par exemple, on trouve des indications syntaxiques en plus grande quantité et de meilleure qualité dans (1a), que dans (1b, 1c) où il manque des informations, ou bien dans (1d, 1e) où figurent des écarts à la grammaire, qu'ils soient volontaires et acceptés (1d) ou non (1e).

(1) a. Deux scientifiques travaillent actuellement sur un matériau qu'ils ont baptisé "claytronics", qui sera composé de nano-machines capables de s'organiser pour reproduire à distance en 3D, et via internet, n'importe quelle forme, y compris la vôtre. (`fr.sci.philo`, 21 septembre 2006)

b. lundi lavage mardi repassage mercredi repos (Mertens, 1993)

c. il y a mon frère sa moto le guidon eh ben complètement naze quoi (Cappeau & Deulofeu, 2001)

d. T'occupe !

e. c'est parcequ'on est 3 à bosser sur le site et que le troisième à casser son portable alors la com passe mal ! (`http://www.onpeutlefaire.com/forum/index.php?showtopic=6164`, 26 juillet 2006)

Notre grammaire est donc conçue comme une ressource unique, rassemblant des constructions provenant de tous les domaines selon une représentation homogène.

Une analyse multi-dimensionnelle. Le dernier point théorique concerne les axes de structuration de l'analyse linguistique. Nous pensons, pour ce qui nous concerne, que la structure arborescente telle que définie par (Chomsky, 1957) (et qui équivaut à la plupart des structures de représentation des grammaires formelles, y compris celles de dépendance (Kahane, 2006)) ne suffit pas. En effet, l'une des premières choses que l'on apprend en linguistique est que l'information y est organisée selon deux axes, syntagmatique et paradigmatique (De Saussure, 1916). Cette idée fondamentale a été suivie dans certaines descriptions (Meillet, 1924; Bally, 1965; Perrot, 1994), mais elle est jusqu'à présent demeurée ignorée dans les grammaires formelles : la dimension paradigmatique, considérée *a priori* comme restant « *virtuelle* » (Dubois *et al.*, 1994), n'est pas prise en compte et ne peut y être représentée. Pourtant il a été montré (Blanche-Benveniste *et al.*, 1990) que des objets produits dans un énoncé peuvent entretenir des

relations s'inscrivant dans cet axe, comme les disfluences et les coordinations. En conséquence, nous avons tenu dans notre modèle à bien considérer l'analyse syntaxique comme s'articulant simultanément selon les deux dimensions syntagmatique et paradigmatique et à permettre l'introduction de constructions des deux sortes.

2.2 Aspects formels

Une grammaire[1] en GP se présente sous la forme d'un couple constitué d'un *ensemble de descriptions linguistiques* et d'une *spécification des types de propriétés*.

Ensemble de descriptions linguistiques. La première partie se présente sous la forme d'un réseau complexe de descriptions de *constructions* linguistiques (Fillmore, 1985), héritant les unes des autres. Chaque objet y est représenté sous la forme d'un double ensemble, de caractéristiques intrinsèques d'une part (*i.e.*, les qualités inhérentes à la construction, p.ex. le fait de porter un genre pour une construction nominale), et extrinsèques d'autre part (*i.e.*, les qualités qui lui viennent de ses constituants, p.ex. l'accord, entre un déterminant et un nom et/ou entre un nom et un adjectif, *etc.*) (Guénot, 2005a). Les informations intrinsèques prennent la forme d'une matrice de traits, et les extrinsèques d'une liste de contraintes : les *propriétés*.

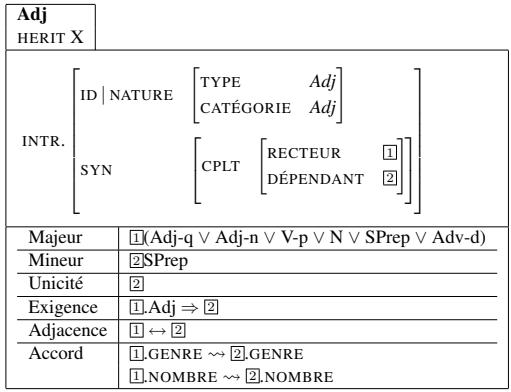


FIG. 1 – Représentation d'une construction en GP : l'Adjectif construit (Adj).

La Figure 1 montre la forme que revêt une construction en GP : il s'agit de l'exemple de la construction adjectivale dans la grammaire qui nous intéresse. Elle est constituée d'un cartouche permettant de l'identifier dans la grammaire, qui indique l'étiquette de la construction (Adj) ainsi que son héritage (ici, Adj hérite de la construction étiquetée X). On y voit ensuite les deux blocs de caractéristiques : en haut la structure de traits rassemblant les informations intrinsèques, indiquant p.ex. le fait qu'il s'agit d'une construction de CATÉGORIE adjectivale ; en bas la liste des propriétés représentant les informations extrinsèques, comme p.ex. l'indication selon laquelle les deux constituants possibles de la construction, s'ils sont présents, doivent être

[1] Ou un module dans le cadre d'un modèle modulaire. P.ex., notre modèle n'a qu'une grammaire, mais une GUST (Kahane, 2002) a autant de grammaires que de modules.

contigüs (quel que soit leur ordre relatif), ou le fait que l'objet Mineur doit s'accorder en genre et en nombre avec le Majeur, si toutefois il possède les traits correspondants.

Spécification des types de propriétés. La seconde partie de la grammaire rassemble la définition des contraintes pouvant s'établir entre les objets de la grammaire, dans un énoncé. A chaque *ensemble de descriptions* correspond une *spécification* donnée, indiquant les relations que la grammaire observe, les objets qu'elles mettent en relation (leur nombre, leur forme, *etc.*), et la façon dont elles sont évaluées (leurs conditions d'évaluation et de satisfaction).

TYPE	DÉFINITION INFORMELLE	EXEMPLE
Majeur	Influe sur la nature	$M(Det)$ dans SN
Mineur	N'influe pas sur la nature	$m(SPrep)$ dans SN
Unicité	Objets uniques	Objet direct dans SV
Exigence	Obligation de cooccurrence	$SPrep \Leftrightarrow N$ dans SN
Exclusion	Restrictions de cooccurrence	$Clitique\ accusatif \not\Leftrightarrow SN\ objet$ dans SV
Précédence	Ordre relatif	$Det \prec Nom$ dans SN
Adjacence	Contiguïté	$Adj \leftrightarrow Nom$ dans SN
Accord	Correspondance de traits	$Nom.genre \rightsquigarrow Adj.genre$ dans SN

FIG. 2 – Spécification des types de propriétés pour le modèle de grammaire présenté.

Par exemple, le modèle que nous avons défini contient dans sa spécification les types de propriétés de la Figure 2 : les deux premières propriétés (Majeur et Mineur) expriment les relations de constituance ; les trois suivantes (Unicité, Exigence, Exclusion) les relations de cooccurrence ; les deux suivantes (Précédence, Adjacence) les relations de position ; enfin, la dernière (Accord) exprime une relation de dépendance.

Il est possible de faire varier cet ensemble : dans d'autres grammaires[2] les types de propriétés ne sont pas les mêmes, et même les propriétés à la dénomination identique n'ont pas nécessairement le même fonctionnement (p.ex. la Précédence peut être définie comme immédiate, ou non).

Il est à noter que GP ne fait pas usage d'informations grammaticales au-delà de la grammaire (*i.e.*, elle n'exprime pas de « principes » comme en HPSG ou en TAG) : avec le jeu des héritages et l'indépendance d'expression des types de propriétés par rapport aux constructions, toutes les informations sont exprimées au sein même de la grammaire et sont traitées selon une même procédure.

3 La grammaire

Le *modèle* introduit dans la section précédente a servi de structure au développement de la *grammaire* que nous présentons dans cette section. Dans son état actuel, celle-ci est constituée d'un ensemble de descriptions de constructions syntaxiques du français. Il ne s'agit pas là de son contenu définitif puisque celle-ci a vocation à être complétée, aussi bien en ce qui concerne la finesse des descriptions syntaxiques que l'ajout d'informations provenant d'autres domaines.

[2] *Cf.* par exemple la grammaire développée pour les besoins de la campagne EASY (Balfourier *et al.*, 2005), ou les extraits de grammaires basées sur d'autres modèles dans (Guénot, 2006a).

Elle constitue toutefois d'ores et déjà une ressource à large couverture pour le français, intégrant notamment des descriptions de phénomènes tels que les SN sans tête nominale (2a), un traitement original des clitiques, les coordinations elliptiques ou de catégories différentes (2b), ou même les disfluences (2c).

(2) a. le vrai, le moi, le boire, les avants, le dessus, des mais, un sans faute, un je ne sais quoi, ...
b. Un pas de plus et tu es mort
c. un jeune homme me l'avait **dans lanché lancé dans** la figure

Plutôt que d'en présenter un extrait nécessairement très limité compte tenu de la place, nous préférons proposer ici une description des grandes originalités de notre ressource. L'ensemble de la grammaire prend la forme d'une hiérarchie de constructions : elle comporte une construction *racine*, dont héritent tous les autres objets. Les constructions qui en héritent peuvent être de deux sortes : *syntagmatique* (§3.1) ou *paradigmatique* (§3.2). L'une des conséquences de cette différenciation est que l'on a pu ainsi proposer une description paradigmatique des disfluences et des coordinations (§3.3), ce qui constitue l'une des plus grandes originalités de notre grammaire formelle.

3.1 Redéfinition des constructions syntagmatiques

Les constructions de la grammaire sont basées sur une redéfinition du *syntagme*, elle-même provenant d'une réflexion critique sur la notion de *tête* : là où la lecture classiquement adoptée de Bloomfield (1961) est que les syntagmes peuvent ou non porter une tête selon qu'il sont endocentriques ou exocentriques (Pollard & Sag, 1994), nous posons pour notre part que tous les syntagmes ont une tête, et que c'est la *portée* de l'influence de cette tête qui est variable en fonction de l'endocentricité ou exocentricité de la construction. En effet le texte de Bloomfield peut être compris de manières assez différentes (Zwicky, 1985) et l'on peut également en conclure que tout syntagme, qu'il soit endocentrique ou exocentrique, porte une tête c'est-à-dire un constituant dont la présence influe directement sur la nature de la construction. Toutefois, cette influence a une portée plus ou moins large suivant les autres constituants de la construction (notamment, suivant leur fonction au sein du syntagme).

Syntagmes endocentriques. Les constructions endocentriques occupent, « *globalement* » comme le disait Bloomfield, le même paradigme que leur tête[3]. Elles portent donc la CATÉGORIE ainsi que le TYPE de leur TÊTE: ainsi, des constructions syntaxiques telles que (3) seront de la même CATÉGORIE (*Adj*) et du même TYPE (*Adj*) que leur TÊTE (resp. *provocateur*, *rouge* et *vert*).

(3) a. ...Thierry était **légèrement provocateur**
b. Maturité du potiron « **Rouge vif** d'Etampes »
c. d'un beau **vert émeraude**

De même, les constructions nominales de (4) et les verbales de (5) sont endocentriques et portent la CATÉGORIE et le TYPE de leur TÊTE (resp. *N* et *V*).

[3]On dit « globalement » parce que parce que pour l'instant on s'accorde à employer la définition de la notion de paradigme de Saussure (1916), tout en concédant qu'à l'observation des corpus on constate qu'elle mérite une redéfinition qui soit plus claire et surtout plus rigoureuse.

(4) a. un **livre ennuyeux**
b. une **tarte maison**
c. la **personne venue me remplacer**

(5) a. l'homme **a pris** une cigarette
b. **il ne mange pas** mais grossit tout de même
c. Pierre **lui avait fait donner** un cadeau

Syntagmes exocentriques. Les constructions exocentriques portent bien le TYPE de leur TÊTE, mais pas sa CATÉGORIE. C'est le cas par exemple des Syntagmes Nominaux (2a, 6) et des Syntagmes Verbaux (7) qui sont de TYPE *N* ou *V* mais de CATÉGORIE *SN* ou *SV*.

(6) a. Le petit chien de ma grand-mère
b. on peut être en désaccord avec **les ceusses qui gèrent l'Usenet français** sans user d'un tel vocabulaire qui n'apporte rigoureusement rien au débat (`fr.usenet.abus.d`, 28 septembre 2000)

(7) a. l'homme **a pris une cigarette**
b. je sais où je dois aller
c. Pierre **lui avait fait donner un cadeau**

Notons que d'après ceci, par exemple, il n'y a pas de Syntagmes Adjectivaux dans la grammaire, puisqu'il n'existe pas en français de syntagmes qui soient de TYPE adjectival mais de CATÉGORIE différente (en d'autres termes, qui soient de TYPE adjectival mais qui occupent un paradigme différent de l'adjectif lexical).

3.2 Introduction des constructions paradigmatiques

La plupart des grammaires admettent (et nous en faisons partie) que certaines constructions n'ont pas de tête ; mais dans le cas habituel, la différence entre constructions à tête et construction sans tête[4] les mène à opposer syntagmes endocentriques (avec tête) et exocentriques (sans tête ou à tête particulière). Or on vient de voir que nous adoptons une position différente puisque nous posons que tous les syntagmes ont bel et bien une (et une seule) tête (régulière). Néanmoins, nous admettons également que toutes les constructions n'ont pas de tête, ce qui signifie que dans la grammaire, toutes les constructions syntaxiques ne sont pas des syntagmes.

En effet, en plus des syntagmes on définit un autre type de construction syntaxique : les *constructions paradigmatiques*. Celles-ci sont caractérisées par le fait que leurs constituants n'entretiennent pas des relations hypotactiques, à la différence des syntagmes (où les relations, quelles qu'elles soient, se font hiérarchiquement entre un objet *recteur* et un *dépendant*), et par conséquent ne portent pas de tête.

3.3 Formalisation des disfluences et traitement original des coordinations

Deux grands types de constructions font partie des paradigmes : les *coordinations* et les *disfluences*. Cette position descriptive est basée sur les travaux de l'Approche Pronominale

[4] Ou alors, suivant les variantes possibles de la définition de la notion de tête, faisant l'objet de classes spéciales telles que les « *têtes faibles* » d'Abeillé (2003) par exemple.

(Blanche-Benveniste, 1987), dont nous avons proposé la description, la représentation formelle et l'intégration à notre grammaire.

Il s'agit ici aussi d'une proposition nouvelle et originale, sous deux aspects. D'une part, les disfluences ne sont quasiment jamais formalisées dans les grammaires actuelles ; et quand elles le sont, alors il n'est pas tenu compte de leurs caractéristiques et de ce qu'il a pu être montré de leur fonctionnement et de leurs apports à l'interprétation, ce qui implique des pertes d'information et un manque de rigueur, ainsi qu'une absence de consensus dans leur traitement (Guénot, 2005b). D'autre part, la fait de ne pas considérer les coordinations comme des syntagmes est un point de vue tout à nouveau en linguistique formelle. Nous avons montré (Guénot, 2006a) l'intérêt de cette approche en en illustrant le fonctionnement simple et efficace sur des exemples variés tels que les coordinations mettant en jeu des coordonnés de catégories différentes (2b) ou des coordonnants non standard (8a), ou encore les formes elliptiques (8b).

(8) a. Il a été habiter à côté de chez Rosalie **que** Rosalie elle savait pas (Deulofeu, 1999)
b. Pierre aime le cinéma et Marie le théâtre

4 Mise à disposition

La grammaire que nous venons de présenter est mise à la disposition publique sur le *Centre de Ressources pour la Description de l'Oral*[5], en tant que ressource pour la description et le traitement des données. Elle fait l'objet d'une licence *Creative Commons* (Paternité - Partage des conditions initiales à l'identique[6]). Elle sera régulièrement mise à jour à chaque étape validée de son développement. Elle est pour l'instant présentée sous une forme textuelle, qui sera prochainement complétée par son équivalent électronique[7].

Autres ressources. La diffusion de cette grammaire est une première étape de mise à disposition de ressources GP pour le TAL ; en effet nous ferons suivre à cette grammaire la diffusion du modèle lui-même (sous la forme d'un squelette de développement de grammaire) afin de permettre, le cas échéant, son utilisation pour le développement d'autres grammaires (p.ex., portant sur d'autres langues et/ou mettant en œuvre d'autres positions théoriques).

Nous ajouterons également à cela la grammaire que nous avons développée dans le cadre de la campagne EASY (Paroubek, 2005), représentée dans les mêmes formats. Celle-ci est également formalisée en GP mais est basée sur un modèle théoriquement éloigné de notre présente proposition (Balfourier *et al.*, 2005).

5 Conclusion

Nous avons présenté ici nous nouvelle ressource pour le TAL : une grammaire du français à large couverture qui s'intéresse à la fois aux phénomènes écrits et oraux. Elle est basée sur

[5] `http://crdo.fr`. Le service étant en cours d'élaboration et la grammaire n'y étant à l'heure actuelle pas encore disponible, elle est pour l'instant téléchargeable à l'adresse `http://mlguenot.googlepages.com/Grammaire.pdf`.

[6] `http://creativecommons.org/licenses/by-sa/2.0/fr/`.

[7] En l'espèce d'un document XML généré à partir du même source LaTeX que celui qui est à l'origine du PDF.

un ensemble de positionnements théoriques qui constitue une nouveauté pour la linguistique formelle (non-générativisme, non-lexicalisation, non-modularité et multi-dimensionnalité), et est représentée sous la forme de *Grammaire de Propriétés*.

Elle a vocation à participer à la mise en convergence de la variété considérable des descriptions linguistiques de corpus et des possibilités formelles, apportant aux premières une validation par leur explicitation et leur régularisation (Chomsky, 1957; Pollard & Sag, 1994), et aux secondes un élargissement des phénomènes pris en considération.

Dans son état actuel, elle est vouée à être affinée et complétée, aussi bien concernant la description des phénomènes syntaxiques (on y affinera progressivement le contenu en enrichissant la hiérarchie jusqu'à proposer des descriptions de figements), que l'ajout d'informations provenant d'autres domaines (sémantique, prosodie, discours, voire descriptions multimodales (Guénot & Bellengier, 2004)). Le modèle sur lequel elle se base a été conçu afin que de tels ajouts soient facilités par la forme même de la grammaire.

Elle est mise à la disposition du public afin d'en permettre la diffusion, l'utilisation, la comparaison, la complétion avec la participation de la communauté. Ceci constitue une première étape dans la mise à disposition d'un ensemble de ressources GP pour le TAL.

Références

ABEILLÉ A. (2003). A lexicalist and construction-based approach to coordinations. In S. MUELLER, Ed., *Proceedings of the HPSG03 Conference*, p. 5–24, Michigan State University, East Lansing: CSLI Online Publications.

BALFOURIER J.-M., BLACHE P., GUÉNOT M.-L. & VANRULLEN T. (2005). Comparaison de trois analyseurs symboliques dans une tâche d'annotation syntaxique. In *Actes de TALN 2005 - Workshop EASY*, Dourdan, France.

BALLY C. (1965). *Linguistique générale et Linguistique française*. Paris: Leroux.

BLACHE P. (2004). Constraints: an operational framework for construction grammars. In *Proceedings of the 3rd International Conference on Construction Grammars (ICCG3)*, Marseille, France.

BLACHE P. (2005). Property grammars: A fully constraint-based theory. In H. CHRISTIANSEN, P. SKADHAUGE & J. VILLADSEN, Eds., *Constraint Satisfaction and Language Processing*. Springer.

BLANCHE-BENVENISTE C. (1987). Syntaxe, choix du lexique et lieux de bafouillage. *DRLAV*, **36-37**, 123–157.

BLANCHE-BENVENISTE C., BILGER M., ROUGET C. & EYNDE K. V. D. (1990). *Le français parlé: Etudes grammaticales*. Sciences du langage. Paris: CNRS Editions.

BLOOMFIELD L. (1961). *Language*. New York: Holt.

CAPPEAU P. & DEULOFEU H.-J. (2001). Partition et topicalisation : *il y en a* "stabilisateur" de sujets et de topiques indéfinis. *Cahiers de praxématique*, **37**.

CHOMSKY N. (1957). *Syntactic Structures*. la Hague: Mouton.

DE SAUSSURE F. (1916). *Cours de linguistique générale*. Paris: Payot. réédition de 1990.

DEULOFEU H.-J. (1999). Probèmes méthodologiques de l'analyse morphosyntaxique de *que* en français contemporain. *Recherches sur le Français Parlé*, **15**.

DEULOFEU H.-J. (2006). Les consécutives construites avec *tellement* ont-elles une syntaxe scalaire ? In *La scalarité: autant de moyens d'expression, autant d'effets de sens*, Bruxelles, Belgique.

DUBOIS J., GESPIN L., GIACOMO M., MARCELLESI C., MARCELLESI J.-B. & MÉVEL J.-P. (1994). *Dictionnaire de linguistique et des sciences du langage*. Paris: Larousse.

FILLMORE C. (1985). Syntactic intrusions and the notion of grammatical construction. *BLS*, **11**, 73–86.

GUÉNOT M.-L. (2005a). Des constructions à l'interface entre lexique et grammaire. In S. KAHANE, Ed., *Journée de l'Atala Interfaces Lexique-Grammaire et lexiques syntaxiques et sémantiques*, Paris.

GUÉNOT M.-L. (2005b). Parsing de l'oral: traiter les disfluences. In *Actes de TALN 2005*, p. 323–332, Dourdan, France.

GUÉNOT M.-L. (2006a). *Eléments de grammaire du français pour une théorie descriptive et formelle de la langue*. Thèse de doctorat, Université de Provence, Aix-Marseille I.

GUÉNOT M.-L. (2006b). La coordination considérée comme un entassement paradigmatique: description, représentation et intégration. *Actes de TALN 2006 - Cahiers du Cental*, **2**(1), 178–187.

GUÉNOT M.-L. & BELLENGIER E. (2004). Quelques principes pour une grammaire multimodale du français. In B. BEL & I. MARLIEN, Eds., *Actes de RECITAL 2004*, p. 51–60, Fès, Maroc.

KAHANE S. (2002). *Grammaire d'unification sens-texte: Vers un modèle mathématique articulé de la langue naturelle*. Document de synthèse de l'habilitation à diriger les recherches, Université Denis Diderot, Paris VII.

KAHANE S. (2006). On the status of phrases in head-driven phrase-structure grammar: Illustration by a totally lexical treatment of extraction. In A. Polguère (ed.). Benjamins. A paraître.

MEILLET A. (1924). *Traité de grammaire comparée des langues indo-européennes*. Paris: Champion.

MERTENS P. (1993). Accentuation, intonation et morphosyntaxe. *Travaux de Linguistique*, **26**.

PAROUBEK P. (2005). EASY: Campagne d'évaluation des analyseurs syntaxiques. In *Actes de TALN 2005 - Workshop EASY*, Dourdan, France.

PERROT J. (1994). Eléments pour une typologie des structures informatives. In *Mémoires de la Société de Linguistique de Paris, tome II: La phrase: Énonciation et information*, p. 13–26. Leuven: Peeters.

POLLARD C. & SAG I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

PULLUM G. & SCHOLZ B. (2001). On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In P. DE GROOTE, G. MORRILL & C. RETORÉ, Eds., *Lecture Notes in Artificial Intelligence - Proceedings of the 4th LACL International Conference*, p. 17–43, Berlin: Springer.

ZWICKY A. (1985). Heads. *Journal of Linguistics*, **21**, 1–29.

Architecture compositionnelle pour les dépendances croisées

Alexandre DIKOVSKY
LINA-FRE CNRS 2729, Université de Nantes
Alexandre.Dikovsky@univ-nantes.fr

Résumé. L'article présente les principes généraux sous-jacent aux grammaires catégorielles de dépendances : une classe de grammaires de types récemment proposée pour une description compositionnelle et uniforme des dépendances continues et discontinues. Ces grammaires très expressives et analysées en temps polynomial, adoptent naturellement l'architecture multimodale et expriment les dépendances croisées illimitées.

Abstract. This article presents the general principles underlying the categorial dependency grammars : a class of type logical grammars recently introduced as a compositional and uniform definition of continuous and discontinuous dependences. These grammars are very expressive, are parsed in a reasonable polynomial time, naturally adopt the multimodal architecture and explain unlimited cross-serial dependencies.

Mots-clés : grammaires catégorielles de dépendances, grammaires multimodales, analyseur syntaxique.

Keywords: categorial dependency grammars, multimodal grammars, syntactic parser.

1 Introduction

L'intérêt principal des grammaires de types logiques dont les grammaires catégorielles (GC) est leur lien direct et transparent avec la sémantique formelle compositionnelle. Ce lien est établi pour une phrase générée à travers l'isomorphisme entre une preuve de correction du choix des types pour les mots dans la phrase et l'expession sémantique extraite de cette preuve. Les relations syntaxiques entre les mots définies par les types sont formaliseés par un calcul logique de types qui n'est pas spécifique à une grammaire mais à une classe de grammaires. On construit ainsi des interfaces simples et élégantes entre la syntaxe et la sémantique à la base de principes plus ou moins universels. Tant que les relations entre les mots (*dépendances*) s'accordent bien avec les relations de précédence (*ordre des mots*), à savoir lorsqu'elles ne dépassent jamais les limites des domaines syntaxiques l ocaux des mots (*dépendances projectives*), les preuves de correction sont isomorphes aux systèmes de constituents des phrases. A ce niveau de représentation syntaxique il est en principe possible de définir les types directement en termes de dépendances. En fait, les premières définitions des grammaires de dépendances (GD) (Gaifman, 1961) ont été similaires à celle des GC classiques (Bar-Hillel, 1953) [1]. Cependant, il existe dans

[1] Or, même à ce niveau, on peut remarquer qu'à la différence des grammaires de types logiques, les GD traitent les modifieurs comme adjoints.

toute langue des dépendances *non bornées* par les domaines locaux (*dépendances non projectives*). Elles sont dues aux formes et aux mots fonctionnels discontinus (comme les particules négatives, les pronoms comparatifs, etc.), ou à l'interférence des éléments des structures extra-syntaxiques telles que la structure communicative (cf. la topicalisation), la co-référence, les relations de portée, etc. ou, au contraire, sont dues au manque en surface, des membres des relations sémantiques (comme c'est le cas de la relativisation ou de l'extraction au cours de la coordination). Pour y faire face les calculs logiques sont complétés par des règles qui, d'un côté, rendent les preuves plus flexibles au détriment du lien direct avec les constituents, e.g. en montant les types (Lambek, 1961; Steedman, 1996), et d'un autre côté, les rendent plus sélectives, e.g. en choisissant les règles structurelles spécifiques en foncion de connecteurs différents (règles *multimodales* dues à Oehrle, Morrill, Moortgat et Hepple (Morrill, 1994; Moortgat, 1997)). Avec ces moyens on peut exprimer les dépendances non bornées tout en gardant l'interprétation sémantique compositionnelle. En même temps, à cause de l'expressivité accrue, la complexité des preuves devient exponentielle, voire pire.

Dans l'article (Dikovsky, 2004), nous avons proposé une nouvelle architecture compositionnelle de types invariables de dépendances (sans montée des types). Elle est établie sur la base de la distinction faite entre les types *neutres* des dépendances projectives, qui sont formalisés par la règle classique d'élimination d'arguments, et les types des dépendances non bornées (*valences*) dotés de *polarisation* et d'*orientation*, qui sont formalisés par une règle appelée **FA** (*first available*) de saturation (*appariement*) des valences. La base psycholinguistique de cette règle est l'hypothèse que les dépendances non bornées sont gérées par les piles dans la mémoire dynamique d'analyse. **FA** sélectionne la plus proche valence polarisée duale dans la direction indiquée. Elle est conforme avec la majorité des dépendances non projectives dans maintes langues. On a élaboré différents calculs de dépendances avec la règle **FA** (Dekhtyar & Dikovsky, 2004; Dekhtyar & Dikovsky, 2007). Les *Grammaires Catégorielles de Dépendances* (CDG) correspondantes s'avèrent expressives. En même temps, elles disposent d'algorithmes d'analyse en temps polynomial. Tout de même, la règle **FA** n'est pas universelle. Par exemple, elle n'est pas adaptée aux dépendances croisées illimitées du hollandais exposées dans (Bresnan *et al.*, 1982). C'est pourquoi, dans cet article nous explorons une autre règle d'appariement **FC** (*first cross*) qui sélectionne la première valence polarisée duale croisée dans la direction indiquée. Ainsi, la structure dynamique de mémoire qui correspond à cette règle est la file d'attente. **FC** explique les dépendances croisées illimitées en termes d'un langage simple de *structures de dépendances* et non en termes du langage de copies, comme d'habitude. A l'instar de grammaires multimodales de types, nous définissons les *CDG multimodales* (mmCDG) où les règles d'appariement sont considérées comme les modes de compositionnalité propres aux dépendances non projectives. Nous montrons que la règle **FC** est aussi efficace que la règle **FA** et nous présentons un algorithme d'analyse syntaxique de ces grammaires en temps polynomial.

2 Grammaires catégorielles de dépendances

Les CDG sont des *grammaires catégorielles* (GC) qui, à la différence des GC classiques, définissent explicitement les relations de dépendance entre les mots dans la phrase et non les relations de dominance entre les constituants. Elles peuvent déterminer les *structures de dépendances* (SD) plus générales que les *arbres de dépendances* (AD). Une SD d'une phrase $w = w_1 \dots w_n$ est un graphe orienté dont les nœux sont les mots $w_1, \dots, w_n$ ordonnés par l'ordre dans w, avec un nœux sélectionné (la *tête*) et dont les arcs sont étiquetés par les noms

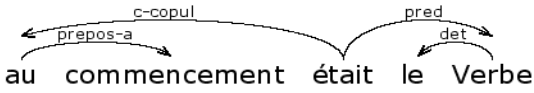


Figure 1

des *dépendances*. E.g., la SD en figure 1 est un AD dont la tête (sa racine) est le mot *était*. Comme toutes les GC, les CDG n'ont pas de règles. Une CDG peut être vue comme un *lexique* qui affecte à chaque mot un ensemble de *types de dépendances*. La particularité essentielle des types des CDG est la distinction faite entre les types de *dépendances projectives* qui relient le gouverneur à ses subordonnés appartenants à son domaine local, et les types de *dépendances non projectives* (*non bornées*) qui le relient aux subordonnés déplacés vers les domaines des autres mots. Les premiers sont définis par les sous types arguments des types du gouverneur, tandis que les derniers sont définis par les *valences* dotées d'une polarisation et d'une orientation (gauche / droite) dont l'ensemble constitue pour chaque type son *potentiel*. Formellement, les types de dépendances sont construits à partir d'un ensemble $\mathbf{C}$ de *types primitifs* et d'un ensemble $V(\mathbf{C})$ de *valences polarisées*. Les éléments de $\mathbf{C}$ sont les noms des relations de dépendance, dont un type sélectionné S (*l'axiome*). Les valences dans $V(\mathbf{C})$ sont orientées : $V(\mathbf{C}) = V^l(\mathbf{C}) \cup V^r(\mathbf{C})$, où $V^l(\mathbf{C})$ consiste des *valences gauches* $\swarrow d$ (*négative*), $\nearrow d$ (*positive*) et $V^r(\mathbf{C})$ consiste des *valences droites* $\nwarrow d$ (positive), $\searrow d$ (négative) où $d \in \mathbf{C}$.
Un *type (de dépendance)* est une expression α^P, où α est un type *basique* et P est un *potentiel*. $gCat(\mathbf{C})$ va noter l'ensemble des types sur $\mathbf{C}$. Les *types basiques* $\mathbf{B}(\mathbf{C})$ sur $\mathbf{C}$ sont les types fonctionnels traditionnels du 1^r ordre destinés à définir les dépendances projectives :

1. $\mathbf{C} \subset \mathbf{B}(\mathbf{C})$. 2. Si $\alpha \in \mathbf{C}$ et $\beta \in \mathbf{B}(\mathbf{C})$, alors $[\alpha\backslash\beta]$, $[\alpha* \backslash\beta]$, $[\beta/\alpha*]$, $[\beta/\alpha] \in \mathbf{B}(\mathbf{C})$. □

Les constructeurs $\backslash, /$ étant supposés associatifs, tout type basique peut être représenté sous la forme $[a_{lm}\backslash ... \backslash a_{l1} \backslash f / a_{r1} / ... / a_{rn}]$. Intuitivement, f est la dépendance du gouverneur et a_{li}, a_{rj} correspondent aux dépendances des subordonnés gauches et droites. $d*$ correspond à la dépendance d itérée. $f = S$ est le type des SD correctes. Les *potentiels* sont les suites de valences polarisées. Ils sont destinés à définir les dépendances non projectives. Dans le cas de dépendances projectives, ils sont vides. Les types avec le potentiel vide sont *neutres*. Par exemple, l'AD projectif en figure 1 est défini par les types neutres suivants :

au $\mapsto [c{-}copul/prepos{-}a]$ *commencement* $\mapsto [prepos{-}a]$ *le* $\mapsto [det]$
était $\mapsto [c{-}copul\backslash S/pred]$ *Verbe* $\mapsto [det\backslash pred]$

Les valences $\swarrow d$ et $\nearrow d$, $d \in \mathbf{C}$, peuvent être vues comme les *parenthèses gauches*. Respectivement, $\nwarrow d$ et $\searrow d$ sont les *parenthèses droites*. Pour une valence gauche, e.g. $\swarrow d$, la valence correspondante (*duale*) droite, $\nwarrow d$, est notée $\check{\swarrow} d = \nwarrow d$. Ensemble ces valences duales appariées définissent la dépendance non projective d. L'adjacence est exprimée en utilisant les types primitifs *d'ancrage* : pour *ancrer* une valence négative $v \in \{\swarrow d, \searrow d \mid d \in \mathbf{C}\}$ (la fin d'une dépendance non projective), c'est-à-dire la placer auprès d'un *mot d'appui*, sont utilisés les types primitifs particuliers *d'ancrage* : $\#(v)$ dont l'élimination signifie l'adjacence des mots et ne crée aucune dépendance. E.g., l'AD non projectif en figure 2 est défini par

Figure 2

les types qui ancrent les clitiques *la, lui* sur l'auxiliaire *a* :

elle $\mapsto [pred]$ $\quad$ *a* $\mapsto [\#(\swarrow clit-iobj)\backslash\#(\swarrow clit-dobj)\backslash pred\backslash S/aux]$
la $\mapsto [\#(\swarrow clit-dobj)]^{\swarrow clit\text{–}dobj}$ $\quad$ *lui* $\mapsto [\#(\swarrow clit-iobj)]^{\swarrow clit\text{–}iobj}$
donnée $\mapsto [aux]^{\nwarrow clit\text{–}iobj \nwarrow clit\text{–}dobj}$

Le sens exact des types est défini par le *calcul de dépendances* suivant [2] :

$\mathbf{L^l}$. $C^{P_1}[C\backslash\beta]^{P_2} \vdash [\beta]^{P_1P_2}$
$\mathbf{I^l}$. $C^{P_1}[C^*\backslash\beta]^{P_2} \vdash [C^*\backslash\beta]^{P_1P_2}$
$\mathbf{\Omega^l}$. $[C^*\backslash\beta]^{P} \vdash [\beta]^{P}$
$\mathbf{D^l_M}$. $\alpha^{P_1(\swarrow C)P(\nwarrow C)P_2} \vdash \alpha^{P_1PP_2}$, si $P_1(\swarrow C)P(\nwarrow C)P_2$ satisfait la règle d'appariement **M**.

$\mathbf{L^l}$ est la règle classique d'élimination. En éliminant le sous-type argument $C \neq \#(\alpha)$, elle crée la dépendance projective C et concatène les potentiels. $C = \#(\alpha)$ ne crée aucune dépendance. $\mathbf{I^l}$ crée $k > 0$ exemplaires de C. $\mathbf{\Omega^l}$ sert pour le cas $k = 0$ et pour éliminer le sous-type itéré. $\mathbf{D^l_M}$ apparie et élimine deux valences duales $\swarrow C$ et $\nwarrow C$ selon la règle d'appariement **M** et crée la dépendance non projective C. Voici deux règles importantes d'appariement :

$\mathbf{FA}^l$: P n'a pas d'occurrence de $\swarrow C, \nwarrow C$ (apparier à la plus proche valence duale disponible).

$\mathbf{FC}^l$: P_1 et P n'ont pas d'occurrences, respectivement, de $\swarrow C$ et de $\nwarrow C$ (apparier à la premiere valence duale croisée, c'est-à-dire à la plus lointaine disponible).

On voit que les valences ressamblent aux traits Slash des GPSG, HPSG, mais à la place de règles complexes de « propagation »des traits Slash les CDG utilisent les règles simples d'appariement **FA** et **FC**. En admettant que toute dépendance non projective C peut avoir sa propre règle d'appariement M_C nous considérons cette règle comme un mode de compositionnalité à travers C. Nous obtenons ainsi par analogie avec l'architecture multimodale pour les grammaires de Lambek (Morrill, 1994; Moortgat, 1997) la notion suivante de grammaire.

Définition 1 *Une* grammaire catégorielle multimodale de dépendances *(mmCDG) est une structure* $G = (W, \mathbf{C}, S, \delta, \mu)$*, où* W *est un vocabulaire,* δ *(le* lexique*) est une fonction qui affecte à chaque mot dans* W *un sous ensemble fini de types dans gCat(***C***) et* μ *est une fonction qui affecte une règle d'appariement à toute dépendance non projective dans* $\mathbf{C}$.
Le calcul de dépendances détermine la relation de prouvabilité correspondante $\vdash_\mu$ *sur les suites de types. La prouvabilité sans règles* **D** *(c'est-à-dire, au cas de dépendances projectives) est notée* $\vdash_{\mathbf{c}}$ *. Pour une SD* D *et une phrase* w*, la relation* $G(D, w)$ *signifie : « D est créée au cours d'une preuve* $\Gamma \vdash_\mu S$ *pour une suite de types* $\Gamma \in \delta(w)$ *».*
Le langage *et le* langage des SD *générés par G sont respectivement les ensembles* $L(G) =_{df} \{w \mid \exists D\ G(D, w)\}$ *et* $\Delta(G) =_{df} \{D \mid \exists w\ G(D, w)\}$*.* $mmCDG^\mu$ *et* $\mathcal{L}(mmCDG^\mu)$ *sont respectivement la famille des grammaires et des langages correspondants.*

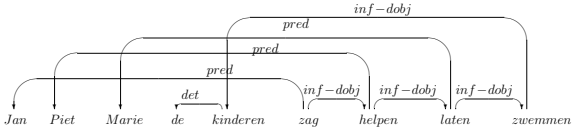
3 Expressivité des mmCDG

Les mmCDG sont très expressives. Avec la règle **FA** elles génèrent tous les langages non contextuels (algébriques), mais aussi maints langages contextuels dont $\{a^n b^n c^n \mid n > 0\}$, les langages $L^{(m)} = \{a_1^n a_2^n ... a_m^n \mid n \geq 1\}$ (Dikovsky, 2004) qui sont faiblement contextuels mais non-TAG à partir de $m > 4$, le langage MIX, qui contient toutes permutations des motifs $a^n b^n c^n, n > 0$, $MIX = \{w \in \{a, b, c\}^+ \mid |w|_a = |w|_b = |w|_c\}$. Or, selon l'hypothèse de E.

[2] Nous exposons les règles gauches. Les règles droites sont symétriques.

Bach, MIX n'est pas faiblement contextuel, ainsi il ne serait pas généré par une grammaire minimaliste, ou multi-TAG, etc. Dans (Dekhtyar & Dikovsky, 2007) on peut trouver d'autres exemples et une preuve du fait que $\mathcal{L}(mmCDG^{\mathbf{FA}})$ est une famille abstraite de langages (AFL).

D'un autre côté, nous croyons (Dikovsky, 2004; Dekhtyar & Dikovsky, 2004) que le langage de copies $L_{copy} = \{ww \mid w \in \{a,b\}^+\}$, qui est généré par une grammaire TAG, n'appartient pas à la famille $\mathcal{L}(mmCDG^{\mathbf{FA},\mathbf{FC}})$. Ce langage est d'un intérêt particulier parce qu'on croit qu'il est un modèle de la construction en néerlandais dite des « *dépendances croisées illimitées* ». Il s'agit des phrases $n_1 n_2 \ldots n_m n_{m+1} v_1 v_{(inf)2} \ldots v_{(inf)m}$, dont un exemple est en figure 3, où il y a une dépendance prédicative $n_1 \overset{pred}{\longleftarrow} v_1$ entre le verbe v_1 en forme finie et le nom n_1, les dépendances prédicatives $n_i \overset{pred}{\longleftarrow} v_{(inf)i}$ entre les verbes $v_{(inf)i}$ à l'infinitif et les noms n_i, pour tout $2 \leq i \leq m$, et éventuellement, une dépendance d'objet direct $n_{m+1} \overset{dobj}{\longleftarrow} v_{(inf)m}$ si le verbe $v_{(inf)m}$ est transitif et le nom n_{m+1} est présent (c'est-à-dire, $n_{m+1} \neq \varepsilon$).



* *Jan Piet Marie les enfants a vu aider faire nager*

Figure 3.

Par ailleurs, une analyse plus approfondie de cette construction (Pulman & Ritchie, 1985) montre que l'accord des formes existe seulement entre n_1 et v_1. Sinon, la forme du nom subordonné est déterminée seulement par le verbe transitif $v_{(inf)m}$ et son argument n_{m+1}. Cela implique que le vrai modèle de cette construction n'est point le langage L_{copy}, mais le langage des SD $\Delta_{cross} = \{D^{(m)} \mid m > 0\}$ sur $W = N \cup V$, où $N \cap V = \emptyset$, $D^{(m)}$ est la SD en figure 4 et $n_{i_l} \in N, v_{j_r} \in V$. En même temps, le *langage* correspondant est algébrique (voire linéaire).

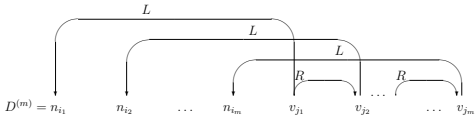


Figure 4. AD $D^{(m)}$

Le langage Δ_{cross} est généré par la $mmCDG^{\mathbf{FC}}$ suivante :

$$G_{cross} = \begin{cases} n & \mapsto & [\#(L)]^{\swarrow_{\mathcal{L}}}, [\#(L)\backslash\#(L)]^{\swarrow_{\mathcal{L}}}, & \text{pour } n \in N \\ v & \mapsto & [\#(L)\backslash S/R]^{\nwarrow\!\!\downarrow}, [R/R]^{\nwarrow\!\!\downarrow}, [R]^{\nwarrow\!\!\downarrow}, & \text{pour } v \in V \end{cases}$$

E.g., une preuve de $D^{(3)} \in \Delta_{cross}$ est montrée en figure 5.

$$\dfrac{\dfrac{\dfrac{[\#(L)]^{\swarrow_{\mathcal{L}}}[\#(L)\backslash\#(L)]^{\swarrow_{\mathcal{L}}}}{[\#(L)]^{\swarrow_{\mathcal{L}}\swarrow_{\mathcal{L}}}}(\mathbf{L}^l) \quad [\#(L)\backslash\#(L)]^{\swarrow_{\mathcal{L}}}}{[\#(L)]^{\swarrow_{\mathcal{L}}\swarrow_{\mathcal{L}}\swarrow_{\mathcal{L}}}}(\mathbf{L}^l) \quad \dfrac{[\#(L)\backslash S/R]^{\nwarrow\!\!\downarrow} \quad \dfrac{[R/R]^{\nwarrow\!\!\downarrow}[R]^{\nwarrow\!\!\downarrow}}{[R]^{\nwarrow\!\!\downarrow\nwarrow\!\!\downarrow}}(\mathbf{L}^r)}{[\#(L)\backslash S]^{\nwarrow\!\!\downarrow\nwarrow\!\!\downarrow\nwarrow\!\!\downarrow}}(\mathbf{L}^r)}{[S]^{\swarrow_{\mathcal{L}}\swarrow_{\mathcal{L}}\swarrow_{\mathcal{L}}\nwarrow\!\!\downarrow\nwarrow\!\!\downarrow\nwarrow\!\!\downarrow}}(\mathbf{L}^l)}{[S]}(\mathbf{D}^l_{\mathbf{FC}} \times 3)$$

Figure 5.

4 Fondements théoriques

Notre solution du problème des dépendances croisées repose sur l'indépendance des types basiques et des valences polarisées dans les preuves du calcul de dépendances. Cette propriété est exprimée en termes de *projections* et de suites de catégories *bien appariées*.

Pour une suite de catégories $\gamma \in gCat(\mathbf{C})^*$ ses projections *locale* $\|\gamma\|_l$ et *de valences* $\|\gamma\|_v$ sont définies ainsi : pour tous $\alpha \in gCat(\mathbf{C})$, $\gamma \in gCat(\mathbf{C})^*$ et $C^P \in gCAT(\mathbf{C})$,

1. $\|\varepsilon\|_l = \|\varepsilon\|_v = \varepsilon$; $\|\alpha\gamma\|_l = \|\alpha\|_l\|\gamma\|_l$ et $\|\alpha\gamma\|_v = \|\alpha\|_v\|\gamma\|_v$
2. $\|C^P\|_l = C$ et $\|C^P\|_v = P$.

Pour un potentiel P, sa projection $\|P\|_d$ sur une paire de valences duales $vd, \check{vd}$ est définie comme $h(P)$ pour l'homomorphisme $h(\alpha) = \alpha$ si $\alpha \in \{vd, \check{vd}\}$ et $h(\alpha) = \varepsilon$ sinon. P est dit *équilibré* si toute projection $\|P\|_d$ est bien appariée au sens habituel.

Soit $|P|_x$ le nombre d'occurrences de x dans P. Alors l'équilibre d'un potentiel P est incrémentalement vérifiable en utilisant les quantités suivantes pour toute $\alpha \in V^l(\mathbf{C})$ et $\check{\alpha} \in V^r(\mathbf{C})$:

$$\Delta_\alpha(P) = max\{|P'|_\alpha - |P'|_{\check{\alpha}} \mid P' \; est \; un \; \text{suffixe} \; de \; P\},$$
$$\Delta_{\check{\alpha}}(P) = max\{|P'|_{\check{\alpha}} - |P'|_\alpha \mid P' \; est \; un \; \text{préfixe} \; de \; P\}.$$

Elles expriment respectivement le *déficit* des $\alpha-$parenthèses droites et gauches dans P (c'est-à-dire, le nombre de parenthèses droites (gauches) qu'il faut rajouter à P de droite (de gauche) pour qu'il devienne équilibré. Les propriétés suivantes sont vérifiées (Dekhtyar & Dikovsky, 2004; Dekhtyar & Dikovsky, 2007) :

Lemme 1 *1. Quels que soient des potentiels P_1, P_2 et des valences $\alpha \in V^l(\mathbf{C})$, $\check{\alpha} \in V^r(\mathbf{C})$,*

$$\Delta_\alpha(P_1P_2) = \Delta_\alpha(P_2) + max\{\Delta_\alpha(P_1) - \Delta_{\check{\alpha}}(P_2), 0\},$$
$$\Delta_{\check{\alpha}}(P_1P_2) = \Delta_{\check{\alpha}}(P_1) + max\{\Delta_{\check{\alpha}}(P_2) - \Delta_\alpha(P_1), 0\}.$$

2. Un potentiel P est équilibré ssi $\sum_{\alpha \in V(\mathbf{C})} \Delta_\alpha(P) = 0$.

La propriété suivante d'*indépendance des projections* (Dekhtyar & Dikovsky, 2004; Dekhtyar & Dikovsky, 2007) garantit l'existence d'un algorithme polynomial d'analyse de $mmCDG^{\mathbf{FA}}$.

Théorème 1 *Pour une mmCDG $G = (W, \mathbf{C}, S, \delta, \mu)$ avec le mode $\mathbf{FA}$ et $x \in W^+$, $x \in L(G)$ ssi il y a une suite $\Gamma \in \delta(x)$ telle que $\|\Gamma\|_l \vdash_{\mathbf{c}} S$ et $\|\Gamma\|_v$ est équilibré.*

Le seul point de sa preuve sensible aux modes est la proposition suivante vraie pour **FA** :

Lemme 2 *Un potentiel P est équilibré ssi pour toute catégorie α^P il y a une preuve $\alpha^P \vdash \alpha$ utilisant exclusivement les règles $\mathbf{D}^l_{\mathbf{M}}$ et $\mathbf{D}^r_{\mathbf{M}}$.*

Pour garantir l'indépendance des projections (et par conséquent, une analyse polynomiale) pour une $mmCDG^M$, il faut prouver ce lemme pour tout mode $\mathbf{M} \in M$. En prouvant le lemme 2 pour $\mathbf{FC}$, nous avons étendu le théorème 1 aux mmCDG avec les modes $\mathbf{FA}, \mathbf{FC}$:

Théorème 2 *Pour $x \in W^+$ et pour une $mmCDG^M$ $G = (W, \mathbf{C}, S, \delta, \mu)$ avec $M = \{\mathbf{FA}\}$, ou $M = \{\mathbf{FC}\}$ ou $M = \{\mathbf{FA}, \mathbf{FC}\}$, $x \in L(G)$ ssi il y a une suite $\Gamma \in \delta(x)$ telle que $\|\Gamma\|_l \vdash_{\mathbf{c}} S$ et $\|\Gamma\|_v$ est équilibré.*

Corollaire 1 $\mathcal{L}(mmCDG^{\mathbf{FA}}) = \mathcal{L}(mmCDG^{\mathbf{FC}}) = \mathcal{L}(mmCDG^{\mathbf{FA},\mathbf{FC}})$.

5 Analyse syntaxique, complexité

Dans l'article (Dekhtyar & Dikovsky, 2004) un algorithme d'analyse en temps polynomial a été décrit pour une version sous commutative du calcul de dépendances [3]. Dans l'article (Dekhtyar & Dikovsky, 2007) cet algorithme a été étendu aux $mmCDG^{\mathbf{FA}}$. Ce même algorithme à un détail près s'applique aussi aux $mmCDG^{\mathbf{FA},\mathbf{FC}}$. Nous l'exposons en figure 6.

Fonctions d'échec. Soit une $mmCDG^M$ $G = (W, \mathbf{C}, S, \delta, \mu)$ avec les valences polarisées gauches $V^l(\mathbf{C}) = \{v_1, \ldots, v_p\}$ et droites $V^r(\mathbf{C}) = \{\breve{v}_1, \ldots, \breve{v}_p\}$. Nous allons d'abord définir deux fonctions d'échec qui vont servir pour une optimisation de l'analyse. Soit $w = w_1 w_2 ... w_n \in W^+$. Alors, pour $1 \leq i \leq n$, $\alpha \in V^l(\mathbf{C})$ et $\beta \in V^r(\mathbf{C})$,

$$\pi^L(\alpha, i) = max\{\Delta_\alpha(\|\Gamma\|_v) \mid \Gamma \in \delta(w_1 ... w_i)\},$$
$$\pi^R(\beta, i) = max\{\Delta_\beta(\|\Gamma\|_v) \mid \Gamma \in \delta(w_{n-i+1} ... w_n)\}$$

sont les *fonction d'échec* gauche et droite. On suppose que $\pi^L(\alpha, 0) = \pi^R(\beta, 0) = 0$.

Algorithme d'analyse syntaxique. mmCdgPars est un algorithme typique de « programmation dynamique ». Il s'applique à une $mmCDG^M$ et à une phrase $w = w_1 w_2 ... w_n \in W^+$ et remplit une matrice triangulaire M dont la dimension est $n \times n$. L'élément $M[i, j], i \leq j$, de M correspond à l'intervalle $w_i ... w_j$ de la phrase et représente un ensemble fini d'« items ». Un *item* est une expression $I = \langle C, \Delta^L, \Delta^R, I^l, I^r \rangle$ qui code une catégorie C^P, où C est une catégorie basique ($C \in \mathbf{B}(\mathbf{C})$), $\Delta^L = (\Delta_{v_1}, \ldots, \Delta_{v_p})$ et $\Delta^R = (\Delta_{\breve{v}_1}, \ldots, \Delta_{\breve{v}_p})$ sont les vecteurs entiers dont chaque composante i correspond à la valence v_i, respectivement $\breve{v}_i$, et vaut le déficit correspondant des v_i-parenthèses droites (gauches) dans le potentiel P. Finalement, I^l, I^r sont les identificateurs des items dans les angles gauches et droites de M à partir desquelles est calculé l'item I (pour tout $I \in M[i, i]$ $I^l = I^r = \emptyset$).

Complexité. Pour une $mmCDG^M$ $G = (W, \mathbf{C}, S, \delta, \mu)$, soit $l_G = |\delta|$ le nombre d'affectations des catégories aux mots dans le lexique, soit $a_G = max\{k \mid \exists x \in W\ ([\alpha_k \backslash ... \backslash \alpha_1 \backslash C / \beta]^P \in \delta(x) \ \vee\ [\beta \backslash C / \alpha_1 / ... / \alpha_k]^P \in \delta(x))\}$ le nombre maximal de sous types arguments dans les catégories affectées, soit $p_G = |V^l(\mathbf{C})| = |V^r(\mathbf{C})|$ le nombre de valences polarisées et $\Delta_G = max\{\Delta_\alpha(P) \mid \exists x \in W(C^P \in \delta(x) \vee \alpha \in V(\mathbf{C}))\}$ le déficit maximal des valences parenthèses dans les catégories affectées. Finalement, soit n la longueur de la phrase analysée.

Théorème 3 *L'algorithme* $\mathbf{mmCdgPars}$ *a une complexité en temps* $\mathbf{O}(\mathsf{l_G} \cdot \mathsf{a_G^2} \cdot (\Delta_\mathsf{G} \cdot \mathsf{n})^{2\mathsf{p_G}} \cdot \mathsf{n}^3)$.

Remarque 1 *1. Pour une grammaire fixée G, les valeurs l_G, a_G, p_G et Δ_G sont constantes. Si G varie, alors le problème d'appartenance devient NP-complèt (Dekhtyar & Dikovsky, 2004).*
2. Si G est sans valence polarisée, alors la complexité est $\mathbf{O}(\mathsf{n}^3)$.
3. Soit le déficit maximal de valences $\sigma_G(n)$ des potentiels survenants dans les preuves des phrases dont la longueur est limitée par n. Si $\sigma_G(n)$ est bornée par une constante c, alors G peut être transformée en une mmCDG G' sans valence polarisée dont le langage est algébrique (Dikovsky, 2001). Or, la taille de G' est exponentielle par rapport à G. Si, de plus, le nombre des dépendances non bornées dans une SD engendrée par G n'est jamais supérieur à une borne constante uniforme (ce qui est typique pour maintes langues), alors la complexité est $\mathbf{O}(\mathsf{n}^3)$ pour la même grammaire G.
4. D'un autre côté, même si **toute** *dépendance de G (sauf S) était définie par une valence polarisée, la complexité serait toujours polynomiale. Cette remarque explique que les mmCDG sont bien adaptées aux langages avec l'ordre flexible. Les limites de cet article ne nous laissent pas faire une analyse plus détaillée de ce cas important.*

[3] L'algorithme a été réalisé en LISP par Darin et Hristian Todorov et en en $C\#$ par Ilya Zaytsev.

Algorithme $\mathbf{mmCdgPars}$
//Entrée : mmCDG G, phrase $w = w_1 ... w_n$
//Sortie : $\langle$"$succès$", DS D$\rangle$ ssi $w \in L(G)$

```
{
  CalcFailFuncL() ;
  CalcFailFuncR() ;
  for (k = 1, ..., n)
  {
      Propose( k )
  }
  for (l = 2, ..., n)
  {
      for (i = 1, ..., n − l)
      {
          j := i + l − 1;
          for (k = i, ..., j − l)
          {
              SubordinateL(i, k, j) ;
              SubordinateR(i, k, j) ;
          }
      }
  }
  if (I = ⟨S, (0, 0, ..., 0), (0, 0, ..., 0), I^l, I^r⟩ ∈ M[1, n])
      return ("succès", Expand(I)⟩ ;
  //procedure Expand( I ) calcule la SD de sortie.
  //Elle seule est sensible aux règles d'appariement
  //FA,FC. Elle est technique et n'est pas incluse
  else
      return ("échec", ∅⟩ ;
}
```

```
//For 1 ≤ i ≤ n
Propose( i )
{
  (loop) foreach (C^P ∈ δ(w_i))
  {
      foreach (v ∈ V^l(C))
      {
          Δ^L[v] := Δ_v(P) ;
          if (Δ^L[v] > π^R[v̈, n − j]) next (loop) ;
          Δ^R[v̈] := Δ_v̈(P) ;
          if (Δ^R[v̈] > π^L[v, i − 1]) next (loop) ;
      }
      AddItem( M[i, i], ⟨C, Δ^L, Δ^R, ∅, ∅⟩ ) ;
  }
}
```

```
AddItem( M[i, j], ⟨C, Δ^L, Δ^R, I^l, I^r⟩ )
{
  M[i, j] := M[i, j] ∪ {⟨C, Δ^L, Δ^R, I^l, I^r⟩} ;
  if (C = [C′ ∗ \β])
  {
      AddItem( M[i, j], ⟨[β], Δ^L, Δ^R, I^l, I^r⟩ ) ;
  }
  if (C = [β/C′∗])
  {
      AddItem( M[i, j], ⟨[β], Δ^L, Δ^R, I^l, I^r⟩ ) ;
  }
}
```

```
CalcFailFuncL()
{
  foreach (v ∈ V^l(C))
  {
      π^L[v, 0] := 0;
      for (i = 1, ..., n)
      {
          π_max := 0;
          foreach (C^P ∈ δ(w_i))
          {
              π_max := max{π_max, Δ_v(P)+
                  max{π^L[v, i − 1] − Δ_v̈(P), 0}};
          }
          π^L[v, i] := π_max;
      }
  }
}
```

CalcFailFuncR() est similaire.

```
//For 1 ≤ i ≤ k ≤ j ≤ n
SubordinateL( i, k, j )
{
  (loop) foreach (I_1 = ⟨α_1, Δ^L_1, Δ^R_1, I^l_1, I^r_1⟩ ∈ M[i, k],
                  I_2 = ⟨α_2, Δ^L_2, Δ^R_2, I^l_2, I^r_2⟩ ∈ M[k + 1, j])
  {
      foreach (v ∈ V^l(C))
      {
          Δ^L[v] := Δ^L_2(v) + max{Δ^L_1(v) − Δ^R_2(v), 0} ;
          if (Δ^L[v] > π^R[v̈, n − j]) next (loop) ;
          Δ^R[v̈] := Δ^R_1(v̈) + max{Δ^R_2(v̈) − Δ^L_1(v̈), 0} ;
          if (Δ^R[v̈] > π^L[v, i − 1]) next (loop) ;
      }
      if ( α_1 = C and α_2 = [C\β] )
      {
          AddItem( M[i, j], ⟨[β], Δ^L, Δ^R, I_1, I_2⟩ ) ;
      }
      elseif ( (α_1 = C and α_2 = [C ∗ \β]) or α_1 = [ε] )
      {
          AddItem( M[i, j], ⟨α_2, Δ^L, Δ^R, I_1, I_2⟩ ) ;
      }
  }
}
```

SubordinateR(i, k, j) est similaire.

Figure 6. Algorithme $\mathbf{mmCdgPars}$

6 Comparaison, discussion

Certes, il y a des grammaires où l'expression des dépendances non bornées ne pose pas problème, e.g. HPSG (Pollard & Sag, 1988), les extensions multimodales des grammaires de Lambek (Morrill, 1994; Moortgat, 1997), dont certaines visent notamment les dépendances (Kruijff, 2001) et leur fournissent une interface compositionnelle avec la sémantique. Or, l'analyse avec ces formalismes expressifs est très complexe et parfois nécessite l'utilisation des systèmes de démonstration des théorèmes. C'est aussi le cas des grammaires qui représentent $PTIME$, dont RCG (Boullier, 2003). A la différence de mmCDG, ces grammaires n'ont pas d'algorithme universel d'analyse en temps $\mathbf{O}(n^k)$, où k dépend de l'alphabet. Cela concerne aussi les grammaires basées sur l'unification et les contraintes, e.g. (Duchier, 1999). Contrairement à ces formalismes, les mmCDG n'utilisent que les moyens primitifs d'une complexité faible. E.g., les Grammaires Topologiques de Dépendances (Duchier & Debusmann, 2001) (voir aussi (Bröker, 1998; Duchier *et al.*, 2004)) utilisent les hierarchies des domaines de l'ordre des mots (WO-domains) qui, en cas de discontinuité, servent à exprimer les contraintes de contiguïté, de distance entre un gouverneur et son modifieur etc. Dans beaucoup des cas, ces contraintes sont exprimées dans mmCDG par le moyen de sous types d'ancrage placés dans les positions correspondantes d'un type du gouverneur.

Les mmCDG représentent une alternative intéressante aux TAG (et équivalentes : CCG, HG (Vijay-Shanker & Weir, 1994)) et aux grammaires faiblement contextuelles (Joshi *et al.*, 1991), telles multi-TAG, non contextuelles multi-composantes, minimalistes, etc. Tout comme ces dernières, les mmCDG disposent d'une analyse syntaxique en temps polynomial. On peut même constater, qu'en pratique l'algorithme **mmCdgPars** va avoir une complexité $\mathbf{O}(n^3)$. Leur avantage décisif est l'architecture compositionnelle de dépendances où toutes les dépendances, projectives comme non bornées, sont définies par les types fonctionnels, ce qui crée la base nécessaire pour une sémantique fonctionnelle de dépendances. En même temps, cette architecture adopte naturellement la multimodalité des dépendances non bornées correspondant aux règles de saturation des valences spécifiques aux différentes langues. Il est important de noter que cette flexibilité syntaxique est atteinte sans explosion du coût de l'analyse syntaxique (par contraste avec les grammaires de Lambek). Malgré leur simplicité, les mmCDG sont très expressives. On a vu que pour exprimer les dépendances croisées illimitées on n'a pas besoin du langage de copies, mais d'un langage des SD facilement exprimé par les mmCDG. Et le fait que MIX est un langage $mmCDG^{\mathbf{FA}}$ montre que ces grammaires sont adaptées aux langues naturelles avec l'ordre des mots flexible.

Enfin, il est difficile de comparer les mmCDG par l'expressivité avec les autres GD qui traitent les dépendances non bornées et qui les analysent en temps polynomial, e.g. (Kahane *et al.*, 1998; Bröker, 2000). Le pouvoir de ces grammaires n'est pas déterminée. Leurs définitions sont opérationnelles (cf. le « lifting »). L'avantage des mmCDG est leur transparence et leur architecture compositionnelle de dépendances.

Références

BAR-HILLEL Y. (1953). A quasi-arithmetical notation for syntactic description. *Language*, **29**(1), 47–58.

BOULLIER P. (2003). Counting with range concatenation grammars. *Theoretical Computer Science*, **293**, 391–416.

BRESNAN J., KAPLAN R., PETERS S. & ZAENEN A. (1982). Cross-serial dependencies in dutch. *Linguistic Inquiry*, **13**(4), 613–635.

BRÖKER N. (1998). Separating surface order and syntactic relations in a dependency grammar. In *Proc. COLING-ACL*, p. 174–180, Montreal.

BRÖKER N. (2000). Unordered and non-projective dependency grammars. *Traitement Automatique des Langues (TAL)*, **41**(1), 245–272.

DEKHTYAR M. & DIKOVSKY A. (2004). Categorial dependency grammars. In M. MOORTGAT & V. PRINCE, Eds., *Proc. of Intern. Conf. on Categorial Grammars*, p. 76–91.

DEKHTYAR M. & DIKOVSKY A. (2007). Generalized categorial dependency grammars. In submission, `www.sciences.univ-nantes.fr/info/perso/permanents/dikovsky/`.

DIKOVSKY A. (2001). Polarized non-projective dependency grammars. In P. DE GROOTE, G. MORILL & C. RETORÉ, Eds., *Proc. of the Fourth Intern. Conf. on Logical Aspects of Computational Linguistics*, volume 2099 of *LNAI*, p. 139–157 : Springer.

DIKOVSKY A. (2004). Dependencies as categories. In *"Recent Advances in Dependency Grammars". COLING'04 Workshop*, p. 90–97.

DUCHIER D. (1999). Axiomatizing dependency parsing using set constraints. In *Sixth Meeting on Mathematics of Language (MOL-6)*, p. 115–126, Orlando, Florida.

DUCHIER D. & DEBUSMANN R. (2001). Topological dependency trees : A constraint-based account of linear precedence. In *Proc. of the Intern. Conf. ACL'2001*, p. 180–187 : ACL & Morgan Kaufman.

DUCHIER D., DEBUSMANN R. & KRUIJFF G.-J. M. (2004). Extensible dependency grammar : A new methodology. In *COLING'04 Workshop*, p. 78–84, Geneva.

GAIFMAN H. (1961). *Dependency systems and phrase structure systems.* Report p-2315, RAND Corp. Santa Monica (CA). Published in Information and Control, 1965, v. 8, n ° 3, pp. 304-337.

JOSHI A. K., SHANKER V. K. & WEIR D. J. (1991). The convergence of mildly context-sensitive grammar formalisms. In P. SELLS, S. SHIEBER & T. WASOW, Eds., *Foundational issues in natural language processing*, p. 31–81, Cambridge, MA : MIT Press.

KAHANE S., NASR A. & RAMBOW O. (1998). Pseudo-projectivity : A polynomially parsable non-projective dependency grammar. In *Proc. COLING-ACL*, p. 646–652, Montreal.

KRUIJFF G.-J. M. (2001). *A Categorial-Modal Logical Architecture of Informativity : Dependency Grammar Logic & Information Structure*. PhD thesis, Charles University, Prague.

LAMBEK J. (1961). On the calculus of syntactic types. In R. JAKOBSON, Ed., *Structure of languages and its mathematical aspects*, p. 166–178. Providence RI : American Mathematical Society.

MOORTGAT M. (1997). Categorial type logics. In J. VAN BENTHEM & A. TER MEULEN, Eds., *Handbook of Logic and Language*, chapter 2, p. 93–177. Elsevier, The MIT Press.

MORRILL G. V. (1994). *Type Logical Grammar. Categorial Logic of Signs*. Kluwer.

POLLARD C. & SAG I. (1988). *An Information Based Approach to Syntax and Semantics, Part I*. Stanford, California : CSLI.

PULMAN S. & RITCHIE G. (1985). Indexed grammars and interesting dependencies. *UEA Papers in Linguistics*, **23**, 21–38.

M. STEEDMAN, Ed. (1996). *Surface Structure and Interpretation.* The MIT Press.

VIJAY-SHANKER K. & WEIR D. (1994). The equivalence of four extensions of context-free grammars. *Mathematical Systems Theory*, **27**, 511–545.

SemTAG, une architecture pour le développement et l'utilisation de grammaires d'arbres adjoints à portée sémantique

Claire GARDENT[1], Yannick PARMENTIER[2]
[1] CNRS / LORIA, Campus scientifique – BP 259
F-54 506 Vandœuvre-Lès-Nancy CEDEX
[2] INRIA / LORIA – Nancy Université, Campus scientifique, BP 259
F-54 506 Vandœuvre-Lès-Nancy CEDEX
{gardent,parmenti}@loria.fr

Résumé. Dans cet article, nous présentons une architecture logicielle libre et ouverte pour le développement de grammaires d'arbres adjoints à portée sémantique. Cette architecture utilise un compilateur de métagrammaires afin de faciliter l'extension et la maintenance de la grammaire, et intègre un module de construction sémantique permettant de vérifier la couverture aussi bien syntaxique que sémantique de la grammaire. Ce module utilise un analyseur syntaxique tabulaire généré automatiquement à partir de la grammaire par le système DyALog. Nous présentons également les résultats de l'évaluation d'une grammaire du français développée au moyen de cette architecture.

Abstract. In this paper, we introduce a free and open software architecture for the development of Tree Adjoining Grammars equipped with semantic information. This architecture uses a metagrammar compiler to facilitate the grammar extension and maintnance, and includes a semantic construction module allowing to check both the syntactic and semantic coverage of the grammar. This module uses a tabular syntactic parser generated automatically from this grammar using the DyALog system. We also give the results of the evaluation of a real-size TAG for French developed using this architecture.

Mots-clés : analyseur syntaxique, grammaires d'arbres adjoints, construction sémantique, architecture logicielle.

Keywords: syntactic parser, tree adjoining grammars, semantic construction, software architecture.

1 Introduction

Un objectif central du traitement automatique des langues est de construire une représentation du sens des textes afin de pouvoir raisonner sur leur contenu. Suivant la granularité de sens désirée, plusieurs approches sont possibles. Typiquement, la recherche d'information s'appuie sur une représentation « à gros grain » où le sens d'un texte est un « sac de mots » (cf.

1a) ; l'extraction d'information demande une représentation plus fine où en particulier les relations sémantiques entre (sens de) constituants doivent être spécifiées (cf. 1b) ; et les systèmes de dialogue, systèmes questions-réponses ou systèmes de détection d'implications textuelles, s'appuient souvent sur une représentation dite « profonde » où des phénomènes tels que la quantification et les modalités pourront être pris en compte (cf. 1c).

(1) L'homme regarde souvent la maison
   a. { *homme, regarde, maison* }
   b. *homme(h), regarde(h,m), maison(m)*
   c. $\exists x \exists y \exists e.$*homme(x)* $\wedge$ *souvent(e)* $\wedge$ *regarde(e,h,m)* $\wedge$ *maison(m)*

Pour construire le troisième type de représentation c.-à-d., une représentation profonde, une approche communément adoptée est de suivre Montague (Montague, 1974) et de développer des grammaires et des lexiques permettant une sémantique compositionnelle c'est-à-dire, une sémantique où le sens d'un constituant est une fonction de la syntaxe de ce constituant et du sens de ses sous-constituants. Ainsi, les grammaires syntagmatiques guidées par les têtes (HPSG, (Copestake *et al.*, 2005)) intègrent une sémantique basée sur les structures à recursion minimale (MRS), les grammaires lexicales fonctionnelles (LFG, (Frank & Van Genabith, 2001)) couplent la construction syntaxique avec une construction sémantique basée sur la sémantique « colle » (glue semantics) et les grammaires catégorielles combinatoires (CCG, (Bos *et al.*, 2004)) utilisent l'isomorphisme de Curry-Howard pour associer de façon systématique, constituants syntaxiques et termes lambda. Pour chacune de ces grammaires, une implantation existe qui démontre la faisabilité de l'approche théorique sous-jacente et en permet l'utilisation pratique dans des systèmes de TAL.

Une exception notoire concerne la construction sémantique dans les grammaires d'arbres adjoints (Joshi *et al.*, 1975). Pour ces grammaires en effet, des propositions théoriques existent mais aucune implantation. Dans cet article, nous reprenons la proposition théorique avancée par (Gardent & Kallmeyer, 2003) et décrivons sa mise en oeuvre dans un système implanté. Nous présentons les différentes composantes du système (grammaire, compilateur de grammaire, module de construction sémantique) et donnons les résultats d'une première évaluation sur une grammaire noyau du français. Utilisé pour développer une grammaire d'arbres adjoints à dimension sémantique pour le français, ce système est à notre connaissance, le premier système logiciel libre permettant la construction de représentations sémantiques profondes pour le français. En effet, il existe une grammaire HPSG pour le français (Tseng, 2003) mais sa couverture est limitée. Une grammaire LFG existe également mais étant développée par Xerox, elle n'est pas disponible pour la recherche. Par contraste, SEMTAG est un logiciel libre et ouvert. Le logiciel de développement est disponible à l 'url `http://trac.loria.fr/~semtag` avec une grammaire jouet. La grammaire est accessible sur demande et sera rendue disponible prochainement.

L'article est structuré de la façon suivante. Nous commençons (section 2) par présenter le modèle linguistique utilisé c.-à-d., les grammaires d'arbres adjoints, la sémantique plate à trous et l'interface syntaxe/sémantique. Nous présentons ensuite brièvement (section 3) la grammaire du français utilisée et donnons quelques chiffres sur sa couverture actuelle. Dans la section 4, nous présentons le module de construction sémantique. Enfin, la section 5 donne les résultats d'une première évaluation du système en termes de couverture et d'ambiguïté syntaxique et sémantique.

2 Modèle linguistique

Le modèle linguistique inclut une grammaire d'arbres adjoints, un langage de représentation sémantique et une modélisation de l'interface syntaxe/sémantique. Les restrictions d'espace nous empêchant de décrire chacune de ces composantes en détail, nous renvoyons le lecteur aux publications sources pour plus de détails.

Formalisme syntaxique : les grammaires d'arbres adjoints (TAG) Les grammaires d'arbres adjoints (Tree Adjoining Grammars, TAG) (Joshi *et al.*, 1975) appartiennent à la famille des grammaires légèrement sensibles au contexte. Une TAG est un système de réécriture d'arbres composé de deux ensembles d'arbres (*arbres initiaux* et *arbres auxiliaires*) et de deux opérations de réécriture (*substitution* et *adjonction*).

Un arbre initial est un arbre dont les noeuds feuilles sont soit étiquetés par des mots, soit des noeuds de substitution (marqués $\downarrow$) c.-à-d., des noeuds où une substitution *doit* prendre place. Un arbre auxiliaire est un arbre contenant un noeud pied (marqué $\star$) – ce noeud pied doit être étiqueté avec la même catégorie que le noeud racine.

Dans la version de TAG que nous utilisons, à savoir les grammaires d'arbres adjoints lexicalisées à structures de traits (FLTAG, (Vijay-Shanker & Joshi, 1988)), les arbres élémentaires sont lexicalisés, c'est-à-dire que pour chaque arbre, au moins un terminal est un lemme ou une forme fléchie. En outre, les noeuds des arbres sont étiquetés par deux structures de traits appelées TOP et BOTTOM. En fin de dérivation, les traits TOP et BOTTOM de chaque noeud sont unifiés.

L'opération de substitution permet d'insérer un arbre élémentaire ou dérivé τ_δ à la frontière d'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 des nœuds en question sont unifiés ($Top_{\tau_\alpha} = Top_{\tau_\delta}$). 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 respectivement avec les traits TOP du noeud racine de l'arbre auxiliaire et les traits BOTTOM de son noeud pied ($Top_n = Top_{Root_{\tau_\beta}}$ et $Bottom_n = Bottom_{Foot_{\tau_\beta}}$).

Formalisme sémantique : la sémantique plate à trous (Hole Semantics) Comme la MRS mentionnée en section 1, le formalisme des sémantiques plates à trous (Bos, 1995) se caractérise par deux points importants. Premièrement, le formalisme permet de sous-spécifier les ambiguïtés de portée – ainsi les interprétations multiples dues à ces ambiguïtés peuvent être représentées de façon compacte. Deuxièmement, (Copestake *et al.*, 2005) ont montré que la structure non récursive des formules plates facilite la réalisation sémantique c.-à-d., la procédure qui permet de produire, à partir d'une représentation sémantique donnée, l'ensemble des phrases associées par la grammaire à cette sémantique. C'est là un point important puisque de fait, la grammaire présentée ici est également utilisée pour la réalisation.

Très brièvement (cf. (Gardent & Kallmeyer, 2003) pour plus de détails), le langage de représentation sémantique L_U utilisé est une reformulation de la logique PLU (Bos, 1995) qui inclut des variables d'unification. Soit I_{var} un ensemble de variables d'unification et I_{con} un ensemble de constantes. Soit H un ensemble de constantes « trous », L_{con}, un ensemble de constantes « étiquettes » et L_{var} un ensemble de variables d'étiquettes ; soit R un ensemble de relations n-aires sur $I_{var} \cup I_{con} \cup H$; et soit $\geq$ une relation sur $H \cup L_{con}$ nommée « a-portée-sur ». Alors,

la syntaxe de L_U est la suivante :

Etant donnés $l \in L_{var} \cup L_{con}$, $h \in H$, $i_1, \ldots, i_n \in I_{var} \cup I_{con} \cup H$ et $R^n \in R$. Alors :

1. $l : R^n(i_1, \ldots, i_n)$ est une formule de L_U
2. $h \geq l$ est une formule de L_U
3. ϕ, ψ est une formule de L_U si ψ est une formule de L_U et ϕ est une formule de L_U
4. Rien d'autre n'est une formule de L_U

En d'autres termes : les formules de L_U sont soit des prédications élémentaires, soit des contraintes de portée, soit des conjonctions de formules. Sémantiquement, ces formules décrivent (ont pour modèle) des formules de la logique du premier ordre.

Par exemple, la représentation sémantique de la phrase « Tout yogi a un guru » est :

(2) $l_0 : \forall(x, h_1, h_2), h_1 \geq l_1, l_1 : Yo(x), h_2 \geq l_2, l_2 : A(x, y), l_3 : \exists(y, h_3, h_4), h_3 \geq l_4, l_4 : Gu(y), h_4 \geq l_2$

Cette formule a deux modèles reflétant les deux interprétations possibles de la phrase d'entrée : soit un même guru existe pour tous les yogis, soit plusieurs.

(3) $l_0 : \forall(x, l_1, l_3), l_1 : Yo(x), l_3 : \exists(x, l_4, l_2), l_4 : Gu(y), l_2 : A(x, y)$
$l_3 : \exists(x, l_4, l_0), l_4 : Gu(y), l_0 : \forall(x, l_1, l_2), l_1 : Yo(x), l_2 : A(x, y)$

Interface syntaxe / sémantique. L'interface entre grammaire et sémantique spécifie la correspondance entre constituants syntaxiques et constituants sémantiques. Cette spécification se fait conformément à la proposition de (Gardent & Kallmeyer, 2003). Chaque arbre élémentaire de la grammaire TAG est associé à une formule sémantique plate où des variables d'unification sont utilisées pour représenter les arguments sémantiques. Ces variables d'unification sont partagées avec des variables apparaissant dans les structures de traits étiquetant les nœuds de l'arbre. Lors de la dérivation TAG, les structures de traits des arbres élémentaires sont unifiées (cf. *supra*), ce qui indirectement, entraîne l'unification des arguments sémantiques. La composition sémantique est ainsi prise en charge par l'opération d'unification inhérente au formalisme TAG. A l'issue de la dérivation, la représentation sémantique de l'arbre dérivé est obtenue en prenant la conjonction des formules élémentaires modulo les unifications ayant eu lieu.

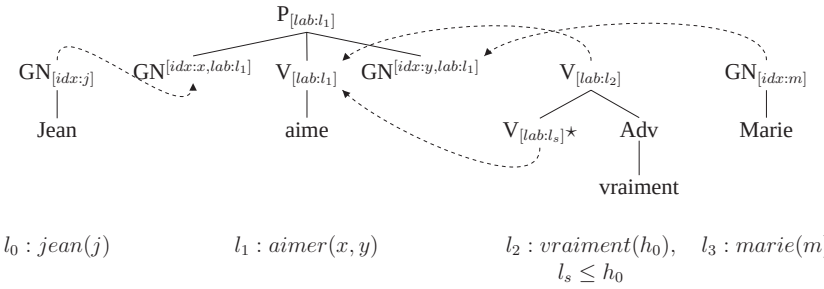


FIG. 1 – Dérivation TAG pour « Jean aime vraiment Marie »

Par exemple, pour la phrase « Jean aime vraiment Marie », la dérivation TAG correspondante est donnée dans la figure 1[1]. Lors de la substitution de l'arbre associé à *Jean* (τ_{Jean}) sur l'arbre associé au prédicat *aimer* (τ_{aimer}), le nœud racine de τ_{Jean} est unifié avec le nœud GN de τ_{aimer} représentant la fonction grammaticale sujet. Le nœud GN de l'arbre résultant contient alors une structure *Top* avec un trait *idx* de valeur x et une structure *Bottom* avec le même trait *idx* ayant la valeur j. A l'issue de la dérivation, les structures *Top* et *Bottom* étant unifiées, la variable x est liée à la constante j. De façon similaire, la variable y est liée à la constante m lors de la substitution de l'arbre τ_{Marie} sur τ_{aimer}. Enfin, l'adjonction de l'adverbe *vraiment* sur le nœud de catégorie V de τ_{aimer} entraîne l'unification de la structure *Bottom* du nœud pied de $\tau_{vraiment}$ avec la structure *Bottom* du nœud d'étiquette V en question, ce qui provoque l'unification de la variable l_s avec la constante l_1. Ainsi, après dérivation et unifications correspondantes, la conjonction des formules sémantiques élémentaires nous donne le résultat escompté, à savoir la représentation sémantique sous-spécifiée suivante :

$$l_0 : jean(j), l_1 : aime(j, m), l_2 : vraiment(h_0), l_1 \leq h_0, l_3 : marie(m)$$

3 Grammaire informatique

La grammaire SEMFRAG est une implantation du modèle linguistique présenté ci-dessus. Spécifiée à l'aide du formalisme XMG (Duchier *et al.*, 2005), cette grammaire est produite par compilation à partir d'une spécification linguistique relativement abstraite et fortement factorisée. La composante syntaxique de la grammaire a été décrite dans (Crabbé, 2005) et la composante sémantique par (Gardent, 2006). Brièvement, l'intégration de l'information sémantique dans une grammaire TAG est facilitée par deux points.

Premièrement, la décoration des arbres élémentaires avec les variables nécessaires à un traitement à grande échelle de la sémantique obéit à un ensemble de principes limités en nombre et relativement rapides à implanter dans le formalisme XMG grâce au haut degré de factorisation permis par ce formalisme. Ces principes sont explicités dans (Gardent, 2007).

Deuxièmement, l'expressivité de XMG facilite la spécification de l'interface syntaxe/sémantique et plus spécifiquement, du partage des variables d'unification entre formules sémantiques et arbres élémentaires. En effet, XMG permet de gérer de manière flexible la portée des variables d'unification manipulées au sein des classes spécifiées par le linguiste. En particulier, ces classes peuvent être associées à des matrices de traits appelées *interfaces* qui sont unifiées lorsque deux fragments sont combinés conjonctivement ou par héritage. Indirectement, cela permet d'unifier des variables introduites dans différentes classes et en particulier, des variables introduites dans des classes syntaxiques (fragments d'arbres) d'une part et dans des classes sémantiques (formules de sémantique plate) d'autre part. Cette fonctionnalité du formalisme nous permet d'encoder de manière relativement aisée l'interface syntaxe / sémantique au niveau métagrammatical, en utilisant la méthodologie suivante :

– chaque fragment d'arbre contenant un nœud lié à une fonction grammaticale représentant un argument sémantique, se voit associé un trait *idx* dont la valeur correspond à une variable partagée avec un trait de l'interface, nommé *FGidx* (où *FG* correspond à la fonction grammaticale en question),

[1] Les structures *top* sont notées en exposants et les structures *bot* en indices. Seuls les traits sémantiques pertinents pour l'exemple sont indiqués.

– chaque foncteur sémantique est associé avec une formule sémantique où les arguments sont des variables partagées avec des traits de l'interface. Ces traits sont nommés en fonction du rôle thématique de l'argument (*p. ex. arg0* ...),
– enfin, dans la règle de combinaison de ces fragments (munie également d'une interface), on ajoute dans l'interface une coindexation entre *FGidx* et l'argument sémantique correspondant (ce qui nous permet également de gérer le cas du passif).

Ce procédé est illustré figure 2, en prenant l'exemple d'un verbe intransitif[2].

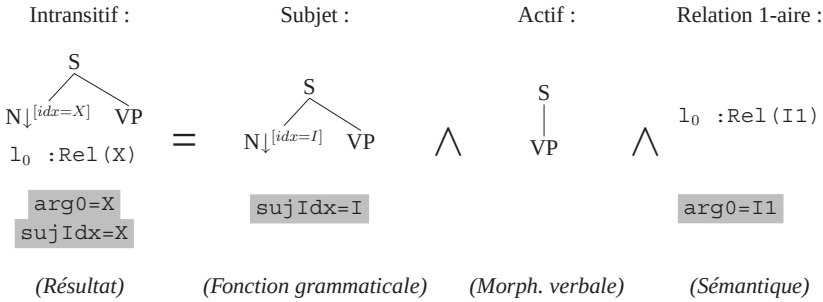


FIG. 2 – Interface syntaxe / sémantique au niveau métagrammatical.

Comme dans le système XTAG, la grammaire SEMFRAG se décompose en 3 sous-modules : un module contenant des d'arbres non lexicalisés groupés en familles[3] ; un lexique de lemmes associant à chaque lemme un prédicat sémantique et une ou plusieurs familles d'arbres ; et un lexique de formes fléchies associant à chaque forme fléchie, un lemme et l'information morpho-syntaxique appropriée. Lors de l'analyse, ces trois modules sont consultés pour associer à chaque mot m de la phrase analysée un arbre lexicalisé (c.-à-d., ancré avec m) dont la sémantique inclut le prédicat spécifié pour m par le lexique de lemme.

4 Construction sémantique

La grammaire SEMFRAG *décrit* l'association entre constituants syntaxiques et représentations sémantiques. Comme le montrent (Gardent & Parmentier, 2005), pour *calculer* cette association (c.-à-d., pour produire la (ou les) représentations sémantique(s) associée(s) par la grammaire à une expression langagière), deux options sont possibles : soit la construction sémantique est intégrée dans l'analyse syntaxique (la construction sémantique se fait pendant la dérivation), soit elle se fait après la dérivation sur la base de cette dérivation et d'un lexique sémantique produit à partir de la grammaire et d'un lexique.

SEMTAG implante la deuxième option, ce qui permet à la fois de rester dans le formalisme TAG (*cf* (Kallmeyer & Romero, 2004)) et de garder une approche modulaire où analyse syntaxique

[2]On remarque que la fonction grammaticale pourrait très bien correspondre à une disjonction des différentes réalisations syntaxiques.

[3]En TAG, une famille d'arbres regroupe tous les arbres élémentaires correspondant à un cadre de sous catégorisation donné *p. ex.*, intransitive.

et construction sémantique restent indépendants l'un de l'autre[4]. Concrètement, le procédé de construction sémantique repose sur le schéma suivant.

Etape 1. Dans un premier temps, toute l'information sémantique incluse dans la grammaire est extraite et stockée dans un lexique sémantique. Ce lexique est en quelque sorte le parallèle sémantique de la grammaire syntaxique TAG au sens où il associe à chaque arbre élémentaire TAG un arbre sémantique correspondant. La figure 3 illustre ce procédé d'extraction pour l'arbre associé à la forme fléchie « dort » (arbre pour une utilisation avec un sujet nominal canonique). L'arbre du haut est celui produit par la compilation de SEMFRAG (suivie de la phase d'ancrage des schémas d'arbres par l'information contenue dans les lexiques de lemmes et de formes fléchies), l'arbre en bas à gauche est l'arbre purement syntaxique extrait de cet arbre et l'arbre en bas à droite, l'arbre sémantique (entrée du lexique sémantique)[5].

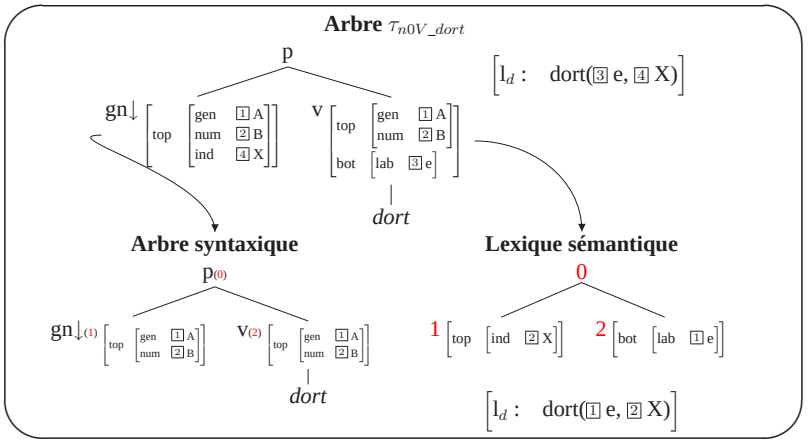


FIG. 3 – Entrée du lexique sémantique.

Etape 2. La deuxième étape consiste à faire une analyse syntaxique de la phrase d'entrée en utilisant uniquement la partie syntaxique de SEMFRAG. Cette analyse est réalisée au moyen du système DyALog (Villemonte de la Clergerie, 2005), un compilateur de programmes logiques avec tabulation des calculs intermédiaires qui permet en particulier de compiler un analyseur syntaxique à partir d'une grammaire TAG donnée. L'analyseur résultant de cette compilation prend en entrée une chaîne préalablement segmentée, et retourne une forêt de dérivation décrivant de façon compacte l'ensemble des dérivations couvrant la chaîne d'entrée. Par exemple, la forêt de dérivation pour la phrase ambiguë « Jean regarde Anne avec le télescope » est celle donnée en figure 4. Cette forêt représente les deux dérivations possibles de la façon suivante. Les nœuds de l'arbres sont étiquetés avec les noms des arbres élémentaires mis en jeu dans la

[4]Cette implantation correspond à la proposition de (Kallmeyer & Romero, 2004), l'avantage étant que les structures étiquetant les nœuds ne contiennent pas de traits à valeur en nombre théoriquement infini (*p. ex.*, les variables de label de la sémantique plate).

[5]Le lexique sémantique est donc calculé par rapport aux arbres ancrés lors de l'analyse syntaxique.

dérivation tandis que les arcs indiquent soit une substitution (trait plein), soit une adjonction (trait en pointillés). Plus précisément, une flèche étiquetée avec l'information $\langle O, n\rangle$ et allant du nœud étiqueté X vers le nœud étiqueté Y, indique que l'arbre X a été combiné par l'opération O ($O \in \{s, a\}$ avec s pour substitution et a pour adjonction) avec l'arbre Y en son nœud n.

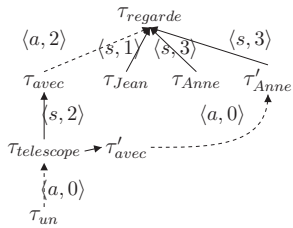


FIG. 4 – Forêt de dérivation de la phrase « Jean regarde Anne avec un télescope ».

Etape 3. Enfin la troisième étape consiste à produire à partir de la forêt de dérivation produite par DYALOG et du lexique sémantique extrait de SEMFRAG, la représentation sémantique de la phrase d'entrée. Pour ce faire, nous avons défini et implanté en Prolog un algorithme de construction sémantique qui traverse la forêt de dérivation dans un processus descendant, et réalise les unifications entre indices sémantiques comme résumé ci-dessous.

On note *Lex(x)* = (τ_x, ϕ_x) l'association spécifiée par le lexique sémantique entre un nom d'arbre syntaxique x, l'arbre sémantique τ_x et la représentation sémantique ϕ_x : *Lex*^{1}*(x)* = τ_x et *Lex*^{2}*(x)* = ϕ_x.
Etant donnée une forêt de dérivation et un lexique sémantique *Lex*, pour construire la (ou les) représentations sémantiques associées, FAIRE :

1. (Initialisation) Pour chaque racine(s) a de la forêt de dérivation, extraire *Lex(a)* = (τ_a, ϕ_a) du lexique sémantique *Lex*. Initialiser la sémantique de a à ϕ_a.
2. (Parcours descendant de la forêt) Pour chaque arc de dérivation de la forme $a_i \overset{o,n}{\rightarrow} a_j$ (où a_i, a_j sont des nœuds représentant des arbres élémentaires, o l'opération utilisée et n l'adresse de Gorn du noeud où a lieu l'opération dans l'arbre désigné par a_j), FAIRE :
   – ajouter ϕ_{a_j} (= *Lex*$^2(a_j)$) à la représentation sémantique ϕ_{a_i} de a_i
   – combiner τ_{a_j} (= *Lex*$^1(a_j)$) avec τ_{a_i} (= *Lex*$^1(a_i)$) conformément à l'opération spécifiée par o, n (les unifications correspondantes prennent place cf. section 2 instanciant par « ric hochet » les variables d'unification présentes dans les représentations sémantiques).
3. Lorsque toutes les dérivations ont été traitées, les structures *Top* et *Bottom* étiquetant chacun des nœuds des arbres sémantiques impliqués dans la dérivation sont unifiées.

En résumé : l'algorithme parcourt chaque arbre de la forêt de dérivation ; collecte la sémantique associée par le lexique sémantique avec chaque nœud (c.-à-d., arbre élémentaire) de cet arbre de dérivation ; et utilise les arbres sémantiques associés par le lexique sémantique aux noms d'arbres syntaxiques, pour retranscrire au niveau sémantique, les unifications correspondants aux opérations d'adjonction et de substitution réalisées au niveau syntaxique. Par exemple, pour la phrase « Jean court », l'algorithme procède comme suit :
– Initialisation : ϕ = { *l0 :courir(X)* }

– Traitement de l'arc $\tau_{Jean} \overset{s,1.0}{\rightarrow} \tau_{court}$:
  – Incrément de la sémantique : ϕ = { *l0 :courir(X), l1 :jean(j)* }
  – Effets des unifications dues à la substitution de l'arbre sémantique associé à τ_{Jean} dans l'arbre sémantique associé à τ_{court} : ϕ = { *l0 :courir(j), l1 :jean(j)* }

5 Evaluation

L'évaluation de cette architecture repose sur une évaluation de la grammaire du français qu'elle a permis de développer, en l'occurence la grammaire SEMFRAG présentée précédemment. Cette grammaire décrit 87 familles d'arbres (cadres de sous-catégorisation), les lexiques utilisés contiennent 1 471 formes fléchies, rattachées à 603 lemmes. L'évaluation consiste à vérifier les caractéristiques suivantes :
– la couverture syntaxique et sémantique sur une suite de tests combinant la *Test Suite for Natural Language Processing (TSNLP)* avec une suite de tests complémentaire (SEMTEST)[6],
– le taux moyen d'ambiguïté sémantique (nombre d'analyses sémantiques par phrase).
Développée dans les années 90s, la TSNLP (Lehmann *et al.*, 1996) est une suite de tests visant à permettre l'évaluation et la comparaison d'analyseurs syntaxiques sur un ensemble controlé et annoté de données. Sur un ensemble de 1 495 phrases tests, SEMTAG a actuellement une couverture syntaxique de 62.88 % et une couverture sémantique de 61.27 %. Le taux d'ambiguïté sémantique moyen est de 2.46.

Bien qu'elle ait été pensée pour une évaluation systématique des constructions syntaxiques, la TSNLP échoue à prendre en compte certains types de variations dont en particulier, les variations sur la réalisation des arguments (canonique, relatif, questionné, cliticisé, clivé, etc.), les variations sur la sous-catégorisation des verbes, les variations sur le type de verbe (verbes à contrôle, à montée, semi-auxiliaire, etc). Pour pallier ce manque, nous l'avons complémentée, avec une suite de phrases illustrant ces variations. Pour cette suite complémentaire, la couverture syntaxique est de 86.78 % et la couverture sémantique de 85.02 %. Le taux d'ambiguïté sémantique moyen est de 3.14.

6 Conclusion

SEMTAG permet d'associer à une phrase du français une représentation profonde de sa sémantique compositionnelle. Comme la section précédente l'a montré, la grammaire utilisée est insuffisante pour avoir une couverture large. Pour traiter du passage à échelle, il serait intéressant d'intégrer dans SEMTAG les techniques de fouilles d'erreur et d'analyse à partir d'arbres factorisés utilisées par (Sagot & Villemonte de La Clergerie, 2006). Par ailleurs, il importe d'évaluer la qualité et l'utilité des représentations sémantiques produites soit par le biais d'applications telles que la reconnaissance d'implications textuelles, soit par le biais de la génération (la sémantique produite permet elle de re-générer la phrase de départ ?).

[6] Par couverture, nous entendons la production d'une représentation syntaxique / sémantique validée manuellement dans un premier temps.

Références

BOS J. (1995). Predicate Logic Unplugged. In *Proceedings of the tenth Amsterdam Colloquium, Amsterdam*, p. 133–142.

BOS J., CLARK S., STEEDMAN M., CURRAN J. R. & HOCKENMAIER J. (2004). Wide-coverage semantic representations from a ccg parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, p. 1240–1246, Geneva, Switzerland.

COPESTAKE A., FLICKINGER D., POLLARD C. & SAG I. A. (2005). Minimal Recursion Semantics : An introduction. *Research on Language and Computation*, **3.4**, 281–332.

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. (2005). XMG : Un Compilateur de Métagrammaire Extensible. In *Actes de TALN 2005, Dourdan, France*, p. 13–22.

FRANK A. & VAN GENABITH J. (2001). GlueTag - Linear Logic based Semantics for LTAG – and what it teaches us about LFG and LTAG –. In *Proceedings of LFG01, Hong Kong*, p. 104–126.

GARDENT C. (2006). Intégration d'une dimension sémantique dans les grammaires d'arbres adjoints. In *Actes de la conférence TALN 2006*, p. 149–158.

GARDENT C. (2007). Tree Adjoining Grammar, Semantic Calculi and Labelling Invariants. In *Proceedings of IWCS 7*, p. 75–85.

GARDENT C. & KALLMEYER L. (2003). Semantic construction in FTAG. In *Proceedings of the European chapter of the Association for Computational Linguistics (EACL'03), Budapest*, p. 123–130.

GARDENT C. & PARMENTIER Y. (2005). Large scale semantic construction for tree adjoining grammars. In *Proceedings of LACL05, Bordeaux, France*, p. 131–146.

JOSHI A., LEVY L. & TAKAHASHI M. (1975). Tree adjunct grammars. p. 136–163. Journal of Comput. Syst. Sci., Vol. 10-1.

KALLMEYER L. & ROMERO M. (2004). LTAG Semantics with Semantic Unification. In *Proceedings of TAG+7. Vancouver*, p. 155–162.

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*, p. 711–716, Kopenhagen.

MONTAGUE R. (1974). English as a formal language. *Formal Philosophy. Selected papers of Richard Montague, pages 188-221*.

SAGOT B. & VILLEMONTE DE LA CLERGERIE E. (2006). Error mining in parsing results. In *Proceedings of ACL 2006*, p. 329–336, Sydney, Australia.

TSENG J. (2003). Lkb grammar implementation : French and beyond. In E. B. ET AL, Ed., *Workshop on Ideas and Strategies for Multilingual Grammar Development*, p. 91–97, Technische Universität Wien.

VIJAY-SHANKER K. & JOSHI A. K. (1988). Feature structures based tree adjoining grammars. In *COLING*, p. 714–719.

VILLEMONTE DE LA CLERGERIE E. (2005). DyALog : a tabular logic programming based environment for NLP. In *Proceedings of CSLP'05*, p. 18–33, Barcelona.

Session
Désambiguïsation

Utiliser des classes de sélection distributionnelle pour désambiguïser les adjectifs

Fabienne VENANT
LaLIC – Université Paris IV, Maison de la recherche,
28 rue Serpente 75006 Paris
fabienne.venant@ens.fr

Résumé. La désambiguïsation lexicale présente un intérêt considérable pour un nombre important d'applications, en traitement automatique des langues comme en recherche d'information. Nous proposons un modèle d'un genre nouveau, fondé sur la théorie de la construction dynamique du sens (Victorri et Fuchs, 1996). Ce modèle donne une place centrale à la polysémie et propose une représentation géométrique du sens. Nous présentons ici une application de ce modèle à la désambiguïsation automatique des adjectifs. La méthode utilisée s'appuie sur une pré-désambiguïsation du nom régissant l'adjectif, par le biais de classes de sélection distributionnelle. Elle permet aussi de prendre en compte les positions relatives du nom et de l'adjectif (postpostion ou antéposition) dans le calcul du sens.

Abstract. Automatic word sense disambiguation represents an important issue for many applications, in Natural Language Processing as in Information Retrieval. We propose a new kind of model, within the framework of Dynamical Construction of Meaning (Victorri and Fuchs, 1996). This model gives a central place to polysemy and proposes a geometric representation of meaning. We present here an application of this model to adjective sense disambiguation. The method we used relies on a pre-disambiguation of the noun used with the adjective under study, using distributionnal classes. It can also take into account the changes in the meaning of the adjective, whether it is placed before or after the noun.

Mots-clés : traitement automatique des langues, désambiguïsation, sémantique, polysémie adjectivale, construction dynamique du sens, synonymie, classes distributionnelles, corpus, espace sémantique, espace distributionnel.

Keywords : natural language processing, word sense disambiguation, semantics, adjectival polysemy, dynamical construction of meaning, synonymy, distributionnal classes, corpus, semantic space, distributional space.

1 Enjeux actuels pour la désambiguïsation automatique

Le développement et la généralisation de l'utilisation de documents numériques génèrent des nouveaux besoins en analyse textuelle, notamment pour la navigation dans des bases de données numériques et la recherche d'information. La désambiguïsation automatique constitue

une étape importante dans l'analyse de ces données textuelles, dans les phases d'indexation, ou de description, des documents, comme dans celle de l'extension de requête. Un des enjeux dans ce domaine est, en effet, de pouvoir faire des requêtes non pas sur des occurrences de mots simples mais sur des concepts, et donc de construire des outils capables de prendre en compte la polysémie. Le but du travail présenté ici est donc d'évaluer les performances d'un modèle dynamique de calcul du sens, et des outils géométriques qu'il met en jeu, dans des tâches de désambiguïsation sémantique d'assez haut niveau. La polysémie adjectivale constitue pour cela un champ d'expérimentation idéal, assez peu exploré d'un point de vue informatique. Les travaux existants portent surtout sur la catégorisation des adjectifs, en lien avec l'acquisition lexicale ((Rasking et Nurinburg, 1996) ; (Bouillon et Viegas, 1999)). Les travaux en désambiguïsation automatique se sont quant à eux majoritairement intéressés aux noms et aux verbes. Les phénomènes mis en jeu dans la sémantique adjectivale sont en effet très subtils, difficiles à formaliser et à expliquer de façon systématique. La prise en compte de ces phénomènes peut cependant beaucoup apporter à des outils de recherche d'information, notamment dans des études d'opinions ou de modalités sentimentales.

2 Un modèle dynamique du sens

Le système informatique que nous avons développé met en jeu le modèle dynamique du sens proposé par Victorri et Fuchs (1996). Le principe est le suivant : on associe à chaque mot polysémique un espace sémantique dans lequel se déploient ses différents sens. Les autres mots présents dans l'énoncé définissent une fonction potentielle, et ce sont les sommets de cette fonction potentielle qui permettent de déterminer le sens pris par le mot étudié, dans l'énoncé considéré. Le développement du système s'est fait en deux étapes. Un premier travail, portant sur la représentation du sens, a consisté à mettre au point une méthode de construction automatique des espaces sémantiques et d'exploration du lexique, par le biais de la relation de synonymie ((Ploux et Victorri, 1998) ; (Venant 2007)). Nous présentons cette méthode en section 3. Un second travail a permis d'utiliser ces espaces sémantiques dans une méthode de désambiguïsation automatique. Les travaux ont porté d'une part sur la prise en compte de la polysémie verbale (Jacquet, 2006), d'autre part sur la prise en compte de la polysémie adjectivale (Venant, 2006).

Nous présentons ici les dernières avancées dans la désambiguïsation adjectivale. Les études linguistiques sur l'adjectif s'accordent pour dégager deux caractéristiques principales de la sémantique adjectivale. La première concerne le fait que le sens de l'adjectif dépend du nom qui le régit. Ainsi *sec* prend des sens différents dans *un terrain sec* et *un visage sec.* La seconde concerne l'influence sur le sens de l'adjectif de sa position relativement au nom. Ainsi *un curieux homme* n'est pas nécessairement *un homme curieux*. Nous attendons du système qu'il soit capable de prendre en compte ces caractéristiques dans le calcul du sens d'un adjectif en présence d'un nom donné. Avant de lancer une étude à grande échelle, nous avons voulu étudier la plausibilité du système par une étude en profondeur des adjectifs *sec*, *curieux* et *méchant*. Notre travail s'appuie une pré-désambiguïsation du nom par le biais de classes de sélection distributionnelle. La section 4 présente la méthode de construction automatique de ces classes. Les sections 5 et 6 détaillent la façon dont nous utilisons ces classes, ainsi que les espaces sémantiques, pour prendre en compte l'influence du nom recteur dans le calcul du sens d'un adjectif. La section 7 porte sur le traitement des changements de sens entre anté et postposition.

3 Construction des espaces sémantiques

Nous illustrons ici, sur le cas de l'adjectif *méchant*, la méthode de construction des espaces sémantiques mise au point par Ploux et Victorri (1998). Cette méthode repose sur l'analyse d'un graphe de synonymie (Dictionnaire Electronique des Synonymes, DES : www.crisco.unicaen.fr). Le DES fournit le graphe de synonymie de *méchant* : les sommets du graphe sont *méchant* et tous ses synonymes, et il y a un lien entre deux de ces adjectifs lorsque le DES indique un renvoi synonymique. La Figure**Erreur ! Aucun nom n'a été donné au signet.Erreur ! Aucun nom n'a été donné au signet.** montre un extrait du graphe de synonymie de *méchant*.

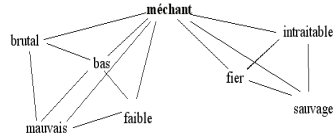


Figure 1 : un extrait du graphe de synonymie de *méchant*

Un synonyme ne suffit pas en général pour définir un sens du mot étudié. On voit, par exemple, sur la Figure **Erreur ! Aucun nom n'a été donné au signet.Erreur ! Aucun nom n'a été donné au signet.** que *bas* est à la fois synonyme de *brutal* et de *faible*, qui correspondent à deux sens différents de *méchant*. On va donc utiliser des ensembles de synonymes, et plus précisément les cliques du graphe. Une clique est un ensemble de sommets deux à deux synonymes le plus grand possible. Le graphe de la Figure **Erreur ! Aucun nom n'a été donné au signet.** présente ainsi 3 cliques : <*bas ; brutal ; mauvais ; méchant*>, <*bas ; faible ; mauvais* ; *méchant*> et <*fier ; intraitable ; méchant ; sauvage*>. On va considérer qu'une clique correspond, en première approximation, à une nuance de sens possible pour le mot considéré. Ce sont donc les cliques qui constitueront les points de l'espace sémantique. L'espace sémantique est alors défini comme l'espace euclidien engendré par *méchant* et tous ses synonymes. A chaque adjectif correspond un axe de l'espace. A chaque clique du graphe correspond un point de l'espace, dont les coordonnées dépendent des synonymes qu'elle contient. Cet espace est muni de la distance du Chi2, bien connue en analyse des données, de façon à rendre compte des proximités sémantiques réelles entre les différents sens du mot étudié. On utilise une Analyse en Composantes Principales pour obtenir une visualisation en deux ou trois dimensions. La Figure **Erreur ! Aucun nom n'a été donné au signet.** présente la visualisation de l'espace sémantique associé à *méchant*. L'espace construit automatiquement rend correctement compte de la sémantique de l'adjectif *méchant*. En effet, on peut constater que les sens de *méchant* se répartissent en trois zones, correspondant aux distinctions de sens apparaissant dans les dictionnaires. En haut à gauche, on trouve les sens intensifs (intensivité négative: *incapable*, *dérisoire, déficient...*), les plus généraux. La partie droite de l'espace sémantique organise les sens les plus spécifiques de *méchant*. En haut à droite, on trouve les cliques correspondant aux sens s'appliquant surtout à des personnes et à leurs actes. En bas à droite on trouve regroupées les cliques correspondant à des sens psychologiques de *méchant*, caractérisant par exemple des attitudes ou des sentiments.

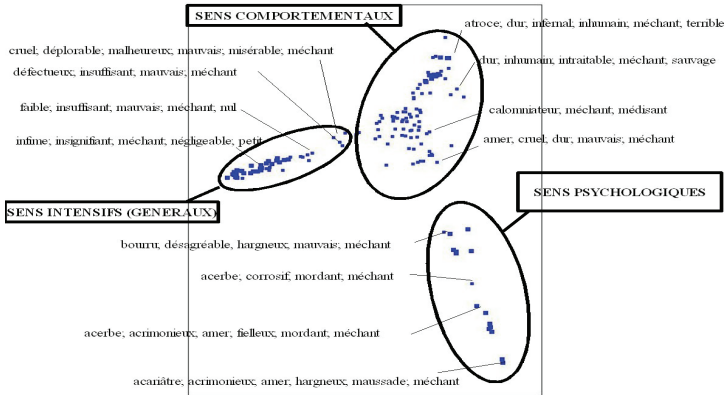
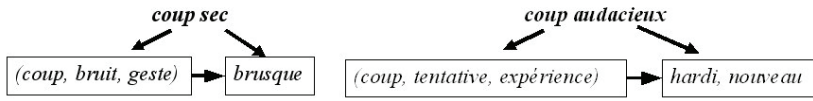


Figure 2 : Espace sémantique associé à *méchant*

4 Utiliser des classes distributionnelles



F

igure 3: Influence mutuelle nom-adjectif

Nous avons, dans un travail antérieur (Jacquet et Venant, 2005), mis au point une méthode de construction automatique de classes de sélection distributionnelle (CSD), à partir d'un corpus. Le principe de construction des CSD est de rassembler dans une même classe les noms influençant de la même manière le sens d'un adjectif. Les noms sont rassemblés sur la base des contextes lexico-syntaxiques qu'ils partagent. Un contexte lexico-syntaxique est constitué d'un mot et d'une relation syntaxique, comme par exemple être recteur de l'adjectif *sec* en position épithète (codé *sec*.EPI), ou être complément d'objet direct du verbe donner (codé *donner*.OBJ). Ces classes vont servir ici à rendre compte de l'aspect dynamique du calcul du sens, en permettant une double désambiguïsation nom-adjectif. Les CSD constituent en effet un outil adéquat pour la prise en charge de l'influence de l'adjectif sur la sémantique du nom qui le régit. Considérons ainsi le groupe nominal *coup sec*. On va ainsi définir une classe (*coup, bruit, geste*..), rassemblant des noms en présence desquels *sec* prend un sens dénotant un manque de douceur. Cette classe se distingue d'une part de celles d'autres noms, comme par exemple celle associée au nom *fruit,* rassemblant des noms (*fruit, haricot, légume*...), en présence desquels *sec* prend un sens dénotant un manque d'eau. Elle se distingue d'autre part de celle associée au même nom *coup*, dans l'étude d'un autre syntagme, par exemple le syntagme *coup audacieux* (*cf.* figure 3). Ce qui nous intéresse particulièrement ici, c'est que la classe d'un nom varie en fonction de l'adjectif étudié, et que c'est cette classe qui permet

ensuite de désambiguïser l'adjectif, et de lui assigner le sens correct en présence du nom considéré.

Pour construire automatiquement ces classes, nous travaillons à partir des sorties de l'analyseur syntaxique Syntex (Bourigault et Fabre, 2000) sur le corpus *Le Monde LM10*. Après filtrage, le corpus contient 31 417 mots et 61 202 contextes lexico-syntaxiques (CLS). A partir de ces données, nous construisons l'espace multidimensionnel engendré par les CLS. C'est ce que nous appelons l'espace distributionnel associé au corpus. Chaque mot y est représenté par un point. Les coordonnées de ce point dépendent des fréquences d'emploi du mot associé dans chacun des CLS engendrant l'espace. La classe d'un nom N en tant que recteur d'un adjectif A rassemble des noms qui sont eux aussi attestés comme étant recteur de l'adjectif A, et qui sont proches du nom N dans l'espace distributionnel. Nous ne détaillons pas ici la méthode de construction de la classe (*cf.* Venant 2006), mais nous l'illustrons sur le cas du nom *bruit.* Pour construire la classe associée au nom *bruit* en tant que recteur de *sec* dans une relation épithète, on commence par chercher dans l'espace distributionnel tous les noms attestés comme recteur de *sec* dans le corpus, c'est-à-dire les noms qui ont une coordonnée non nulle selon la dimension *sec*.EPI. Si cet ensemble contient plus de 100 mots, on ne prend que les 100 mots les plus proches (au sens du Chi2) de *bruit* dans l'espace distributionnel. Notons MOTS l'ensemble formé. On va ensuite recenser tous les contextes pour lesquels au moins un des éléments de MOTS a une coordonnée non nulle, c'est-à-dire l'ensemble des contextes dans lesquels au moins un des noms inclus dans MOTS est employé. Notons CONT l'union de tous ces contextes. Dans le cas de *bruit*, MOTS contient 59 mots (l'ensemble des noms recteurs de *sec* dans le corpus) et CONT contient 9 506 contextes. Une Analyse en Composantes Principales fournit alors les 10 axes de visualisation synthétisant le mieux l'information des 9 506 contextes de CONT, ainsi que les coordonnées des points représentant les 59 mots étudiés dans l'espace euclidien engendré par ces 10 axes. On peut ainsi obtenir des représentations en deux ou trois dimensions de l'espace engendré par CONT. La figure 4, ci- dessous, montre ainsi deux visualisations de l'ensemble des noms recteurs de *sec* dans l'espace distributionnel. On peut ainsi s'apercevoir que *bruit* est proche de *coup* et *geste*, et que *fruit* est proche de *légume* et *pain.* Chaque nom est ainsi proche de noms en présence desquels l'adjectif *sec* prend des sens similaires.

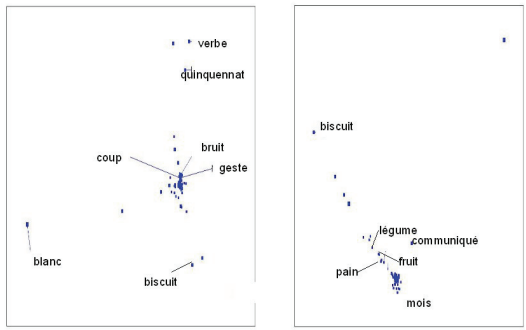


Figure 4: Noms recteur de *sec* dans l'espace distributionnel

Pour constituer la classe distributionnelle de *bruit*, dans le contexte *sec*.EPI, le système classe les noms selon leur distance au nom étudié. Il ajoute ces noms à la classe, du plus proche au

plus éloigné, en additionnant leurs fréquences respectives dans le corpus. Il continue tant qu'un certain seuil de présence dans le corpus n'a pas été atteint. Ce seuil correspond au troisième quartile de la série des fréquences dans le corpus des noms étudiés. Il vaut ici 10 960. La classe distributionnelle de *bruit*, en tant que recteur de l'adjectif épithète *sec*, est (*bruit,* coup).

5 Influence du nom recteur

Nous avons, dans un premier temps, travaillé sur l'influence du nom recteur sur le sens de l'adjectif *sec*. L'influence du nom est prise en compte sous la forme d'une fonction potentielle, définie sur l'espace sémantique de l'adjectif étudié, et dont les sommets permettent de déterminer dans quelle zone de l'espace sémantique ce nom contraint l'adjectif à prendre son sens. Il s'agit donc ici d'associer à chaque nom, ou plutôt à la CSD de ce nom en tant que recteur de *sec*, un potentiel désambiguïsateur. Pour cela, le système que nous avons développé calcule le degré d'affinité de la CSD avec chacune des cliques de *sec,* en fonction des fréquences de cooccurences de chacun des noms présents dans la classe avec chacun des synonymes présents dans la clique. Le tableau 1 présente quelques degrés d'affinités, entre certaines cliques de *sec* et la classe C : (*bruit; coup*).

CLIQUES	Degré d'affinité avec C
austère ; rude ; sec ; simple	98%
bourru ; dur ; rude ; rébarbatif ; sec ; sévère	98%
bourru ; brutal ; dur ; rude ; sec ; sévère ; âpre	97%
sec ; seul ; simple	96%
bourru ; brutal ; cru ; dur ; rude ; sec	95%
bourru ; brusque ; désagréable ; rude ; sec	95%

Tableau 1. Cliques présentant les plus fortes affinités avec la classe C : (bruit, coup)

Ces degrés d'affinité sont ensuite mis en jeu dans le calcul d'une fonction potentielle associée à la classe. La valeur de la fonction en chaque point dépend du degré d'affinité de la classe avec la clique associée à ce point. La fonction potentielle associée à la classe C : (*bruit,coup*) est la suivante : elle atteint son maximum dans la région de l'espace sémantique de *sec* qui rassemble les cliques exprimant le manque de douceur (*ie.* contenant des adjectifs comme *brusque, bref, brutal*...), ce qui correspond bien au sens pris par *sec* dans le syntagme *un bruit sec.*

6 Calcul du sens d'un adjectif en présence d'un nom donné

Nous avons partitionné, manuellement, l'espace sémantique de *sec* en zones correspondant aux sens principaux distingués par les dictionnaires : le manque d'eau et l'improductivité (*une fleur sèche, un terrain sec*), le manque de douceur (*un coup sec*), la maigreur (*un visage sec)*, les sens psychologiques (*un coeur sec*)... La tâche du système consiste ensuite à déterminer automatiquement quelle est la zone de l'espace sémantique correspondant au sens pris par *sec* en présence d'un nom donné. Le système associe une fonction potentielle à chacune de ces zones. Cette fonction dépend des cliques appartenant à la zone. Le système effectue alors un calcul intégral pour comparer ces fonctions aux fonctions potentielles des noms, et déterminer comment se répartit le potentiel désambiguïsateur de chacun des noms étudiés, relativement aux différentes zones de sens. Le tableau 2 donne un aperçu partiel des résultats. La tâche a été menée sur 49 noms. Pour 26 d'entre eux, le système sélectionne la zone de sens adéquate,

pour 16 d'entre eux, le résutat est faux, c'est à dire que la zone sélectionnée n'est pas pertinente. Enfin pour les 7 noms restants, le système reste silencieux et ne sélectionne aucune zone de sens. Il s'agit d'une part des mots *mois, refus, régime* et *vol,* très fréquents dans le corpus, donc seuls dans leur classe distributionnelle*,* mais par ailleurs peu employés avec *sec* et ses synonymes, de sorte que le calcul n'aboutit pas*.* Pour les autre noms, *bois, geste* et *licenciement,* l'apport de la classe reste insuffisant en termes de fréquences de cooccurrence avec les synonymes de *sec*. Pour ces 7 noms, un deuxième calcul a été mené, après inclusion dans la classe distributionnelle du nom le plus proche dans l'espace distributionnel. Le système a pu cette fois sélectionner la zone de sens correcte, à l'exception du nom *mois* pour lequel il sélectionne la zone des sens psychologiques.

NOM	***ZONE DE SENS***	***TAUX D'AFFINITE***
sol	Manque d'eau, improductivité Manque de douceur Maigre, décharné	28% 27% 19%
humour	Sens psychologiques Manque de douceur Manque d'eau, improductivité	25% 22% 18%
cheveu	Manque d'eau, improductivité	77%
ton	Manque de douceur Sens psychologiques	42% 33%
hiver	Sens psychologiques	100%

Tableau 2 : Influence du nom recteur, quelques résultats sur l'adjectif *sec*

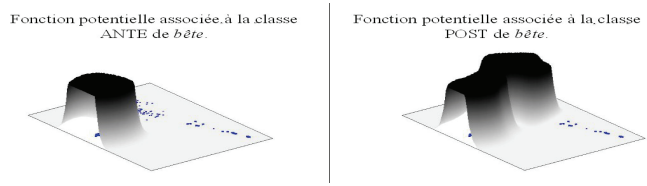
Afin de mesurer l'apport des classes distributionnelles, et donc de la désambiguïsation simultanée du nom et de l'adjectif, ces résultats ont été comparés à ceux obtenus, pour les même noms et sur la même tâche, sans utiliser les classes distributionnelles, c'est-à-dire en associant une fonction potentielle au nom seul, selon ses fréquences d'emplois avec les différents adjectifs.

Il s'avère, et c'est heureux, que l'utilisation des classes distributionnelles ne nous fait perdre aucun des résultats positifs qui apparaissent avec le nom seul. Elle apporte au contraire quelques nuances de sens, puisqu'on voit apparaître le fait qu'un *humour sec*, sec bien sûr d'un point de vue psychologique, est aussi dénué de douceur. De même, les classes distributionnelles permettent au système de détecter que le manque d'eau d'un sol en fait un sol rude, difficile à exploiter, ou encore que *sec* dans *ton sec* est porteur à la fois d'un sens acoustique et d'une connotation psychologique*.* L'utilisation des classes distributionnelles permet aussi d'obtenir des résultats corrects là où l'utilisation du nom seul génère une erreur (c'est le cas pour le nom *arbre*), ou ne donne pas de résultat du tout (14 noms sont concernés dont *bruit* et *cheveu*). Le point important est que l'utilisation des classes distributionnelles ne génère pas d'erreur supplémentaire. Les erreurs obtenues sur des noms comme *hiver* relèvent du mode de description du sens que nous utilisons, et non de l'utilisation des classes distributionnelles. Les points des espaces sémantiques sont en effet des cliques de synonymes. On se heurte ici au fait que non seulement la synonymie entre *froid* et *sec* n'est pas pertinente dans le contexte de *hiver,* mais qu'en plus cela concerne des cliques entières, puisque le phénomène se reproduit pour *glacé* et *glacial*, qui partagent de nombreuses cliques avec *froid* et *sec*. Il faut ici chercher un moyen d'informer notre système que *sec* déploie ses sens dans deux directions sémantiques, l'une plutôt physique, l'autre plutôt psychologique, mais que employé avec certains noms, comme *hiver,* il ne peut prendre son sens que dans le domaine physique, et que donc seules les cliques correspondantes sont à prendre en compte dans le

calcul du sens en présence de ces noms. Une piste pour la résolution de ce problème est l'utilisation d'informations globales sur le lexique adjectival, que nous pensons obtenir automatiquement grâce à notre travail sur la caractérisation des emplois adjectivaux (Venant, 2007).

7 Influence de la position de l'adjectif

Une première étude menée sur l'adjectif *curieux* (François, Victorri et Manguin, 2002) avait montré que le système était capable de rendre compte des changements de sens entre antéposition et postposition. Nous avons poursuivi l'investigation de ce phénomène, par l'étude de l'adjectif *méchant,* dont le sémantisme en antéposition est plus complexe que celui de *curieux*. *Méchant* possède en effet une très grande extension (il peut s'appliquer à n'importe quoi, de la table au costume en passant par l'avocat, la fée, ou la pendule). En antéposition, dans ses emplois généraux, il est sujet au phénomène de désémantisation décrit par Goes (1999), c'est-à-dire qu'il prend un sens si général que le sens du syntagme semble dépendre essentiellement du nom recteur. Enfin, les changements de sens entre antéposition et postposition ne sont pas systématiques. On trouve en effet des cas où le changement de sens est flagrant, et d'autres où le sens de l'adjectif est le même qu'il soit placé avant ou après le nom. Ainsi, s'il est préférable de ne pas engager *un méchant avocat* pour se défendre à un procès, *un avocat méchant* peut au contraire se montrer redoutable. En revanche, on redoutera de la même façon de recevoir *un coup méchant* ou *un méchant coup*. Nous attendons de notre système qu'il soit capable de rendre compte de tous ces phénomènes.
La méthode de désambiguïsation repose sur les mêmes principes que précédemment, mais on utilise ici deux classes distributionnelles [ANTE *vs.* POST] pour chaque mot, l'une calculée à partir des fréquences en antéposition, l'autre calculée à partir des fréquences en postposition. On mesure ensuite l'influence du nom recteur sur la sémantique de *méchant* en associant à chacune de ces classes une fonction potentielle, définie sur l'espace sémantique. Rappelons qu'à chaque point de l'espace sémantique correspond une clique du graphe de synonymie de *méchant*. La valeur de la fonction en chaque point dépend des fréquences de cooccurrence (en antéposition pour une classe [ANTE], en postposition pour une classe [POST]) de chacun des noms constituant la classe avec chacun des adjectifs constituant la clique (pour le détail des calculs voir Venant, 2006). La Figure 5 montre les fonctions potentielles associées aux CSD du nom *bête*.



.Figure 5 : Fonctions potentielles associées aux classes distributionnelles de *bête*

Le système doit alors déterminer le sens que prend l'adjectif *méchant* en présence d'un nom donné. Par sens, nous entendons ici, comme dans le paragraphe précédent, la zone (sens intensifs, sens comportementaux ou sens psychologiques) de l'espace sémantique correspondant au sens de *méchant* dans le syntagme étudié. Ces zones ont été définies manuellement. Le système calcule une fonction potentielle pour chacune des zones. Rappelons que, en chaque point, la valeur de la fonction dépend de l'appartenance ou non de

la clique à la zone considérée. Etant donné un groupe nominal *méchant* + N ou N + *méchant*, le système détermine la CSD [ANTE] ou la CSD [POST], associée à N en tant que recteur de *méchant.* Il calcule ensuite la fonction potentielle associée à cette CSD et la compare à chacune des fonctions potentielles associées aux zones de sens de l'espace sémantique. Les calculs ont été menés pour les 40 noms les plus fréquemment utilisés comme recteur de *méchant* dans le corpus Frantext Catégorisé. Le tableau 3 présente quelques résultats:

ANTE			**POST**	
Sens intensifs Sens comportementaux	76% 24%	cheval	Sens intensifs	69%
Sens comportementaux	100%	regard	Sens comportementaux	100%
Sens intensifs	93%	homme	Sens intensifs	100%
Sens intensifs	99%	loup	Sens comportementaux Sens intensifs	53% 47%

Tableau 3 : Influence de la position de l'adjectif, quelques résultats sur l'adjectif *méchant*

Les résultats que nous obtenons montrent que le système est capable de repérer des changements de sens entre antéposition et postposition. Ce sont les sens intensifs qui obtiennent les scores les plus élevés dans quasiment tous les cas en antéposition. Le système est donc capable de rendre compte du fait qu'on trouve en antéposition les valeurs de sens qui ont la plus grande extension. Ici c'est clairement la valeur intensive, très générale, qui a la plus grande extension. Elle peut s'appliquer à n'importe quoi, alors que les deux autres valeurs ne s'appliquent qu'à des noms animés ou considérés comme tels. Le système repère en outre que, si en antéposition les sens intensifs sont omniprésents, il y a des noms pour lesquels ils ne s'imposent pas forcément. *Cheval, couleur, eau, espèce, farce, matin, mot, nature, parole, taureau, terre* donnent à *méchant* en antéposition tantôt une valeur générale, tantôt une valeur comportementale. Le contexte permet souvent de trancher entre les deux valeurs, mais ce n'est pas toujours très clair. On a ainsi une ambiguïté dans *un méchant cheval*, qui peut désigner selon les cas un cheval maigre, faible ou un cheval agressif. Nous avons pu vérifier que la fonction potentielle associée à la classe [ANTE] de *cheval* en tant que recteur de *méchant* présente deux sommets l'un couvrant la zone intensive et l'autre la zone comportementale. Cette ambiguïté disparaît en postposition, et le système en rend compte. Ainsi c*heval*, *couleur, dent, eau, espèce, farce, maison, mot, nature, parole, part, société, taureau, terre* acceptent aussi bien une valeur intensive que comportementale en antéposition, mais ne sélectionnent que la valeur comportementale ou psychologique en postposition. Pour les noms *bête, bois, bruit, chemin, chose, corps, rire et voix* le changement de sens est encore plus radical, puisqu'en antéposition *méchant* est exclusivement intensif, alors qu'en postposition il devient comportemental ou psychologique. Enfin, le système est aussi capable de repérer les noms pour lesquels on ne repère pas de changement de sens lors du passage de l'antéposition à la postposition. C'est le cas ici de *coup, part, regard* et *vérité*. Les erreurs rencontrées sur des noms comme *homme* ou *enfant* (calcul d'une valeur intensive en postposition) montrent ici encore les limites de la synonymie comme description du sens. Le calcul d'une valeur générale en postposition repose sur les hautes fréquences de cooccurrences de ces noms avec les adjectifs *maigre, faible, pauvre* et *petit.* Or la synonymie entre *méchant* et ces adjectifs n'est plus valable dans le contexte des noms considérés ici, en présence desquels *méchant* se colore plutôt d'une valeur comportementale ou psychologique. Le calcul d'un sens intensif en postposition pour *cheval* montre par ailleurs que ces relations de synonymie, en plus d'être partielles, ne sont valables qu'en antéposition. Là encore, une perspective de résolution du problème repose sur notre méthode d'exploration du graphe adjectival global, qui devrait permettre de repérer automatiquement les sens intensifs d'un adjectif, ceux qui ne sont valables qu'en antéposition.

8 Conclusion

Les analyses détaillées que nous avons menées montrent que les outils informatiques développés sont très prometteurs. Les différentes étapes dans la réalisation, et l'analyse des résultats obtenus à chaque pas, ont mis au jour différents problèmes, que nous devrons résoudre pour aller plus avant. L'utilisation des cliques du graphe de synonymie s'est avérée fort judicieuse, tant pour la construction des espaces sémantiques que pour le calcul du sens proprement dit. Le recouvrement de l'espace sémantique par les cliques contrebalance le fait que la relation de synonymie est une relation partielle, non transitive et peut dépendre de la position de l'adjectif, ce qui cause cependant encore quelques problèmes non résolus. Les classes distributionnelles ont montré leur efficacité pour la prise en compte de l'influence du nom recteur, et de sa position relativement à l'adjectif. Elles constituent une première étape vers une prise en charge des différences entre polysémie nominale et polysémie verbo-adjectivale. Jacquet (2006) dans son travail sur la polysémie verbale, arrive en effet à une conclusion similaire: « On pourrait envisager d'utiliser les CSD pour la désambiguïsation des noms en tant que tels. Cela reviendrait à dire que *bureau* dans l'énoncé *travailler sur le bureau* prend le sens de la classe (*bureau, table, chaise*) que l'on pourrait nommer '*meuble*'. Alors que dans *entrer dans le bureau*, *bureau* prend le sens de la classe (*bureau, cuisine, salon*) que l'on pourrait nommer '*pièce'*. » Le travail décrit ici ne constitue cependant que la première étape dans le processus de validation du modèle utilisé. Nous devons maintenant mener des évaluations massives sur un échantillon beaucoup plus large d'adjectifs ambigus. Il faudra, dans ce cadre, réfléchir au mode de représentation du sens utilisé. On peut chercher à trouver un ensemble de synonymes adéquats pour un adjectif dans un syntagme donné, ou encore réfléchir à une méthode automatique de partitionnement de l'espace sémantique par des outils géométriques (repérer les zones denses en cliques dans l'espace sémantique). Il faudra aussi tenir compte des erreurs inhérentes à l'analyse syntaxique automatique, par exemple dans le répérage du nom recteur de l'adjectif. Ce travail laisse cependant entrevoir tout l'intérêt de l'utilisation des mathématiques du continu en traitement automatique des langues.

Références

BOUILLON P.,VIEGAS E (1999). The description of adjectives for natural languge processing : theoretical and applied perspectives, in *Atelier thématique sur la description des adjectifs pour les traitements informatiques,* Institut d'études scientifiques de Cargëse.

FRANÇOIS J., VICTORRI B. et MANGUIN J.-L. (2002). Polysémie adjectivale et synonymie : l'éventail des sens de curieux, in *La polysémie,* Soutet, Olivier (Eds). Paris : PUS.

FUCHS C., VICTORRI B. (1996). *La polysémie. Construction dynamique du sens*. Hermès.

GOES J. (1999). *L'adjectif. Entre nom et verbe*. Paris/Bruxelles : Duculot.

JACQUET G. (2006), *Polysémie verbale et calcul du sens*; thèse de doctorat de l'EHESS.

JACQUET G., VENANT F. (2005). Construction automatique de classes de sélection distributionnelles, in *Actes du colloque TALN'05*, Dourdan.

PLOUX S., VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement automatique des langues* 39 (1).

RASKIN V., NIRENBURG S. (1996.) Adjectival modification in text meaning representation. Proceedings of *COLING '96.* Copenhagen

VENANT F. (2006), *Représentation et calcul dynamique du sens : exploration du lexique adjectival du français*, thèse de doctorat de l'EHESS.

Disambiguating automatic semantic annotation based on a thesaurus structure

Véronique MALAISÉ[1], Luit GAZENDAM[2], Hennie BRUGMAN[3]
[1] Vrije Universiteit, Amsterdam
[2] Telematica Institute, Enschedé, Netherlands
[3] Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
vmalaise@few.vu.nl, Luit.Gazendam@telin.nl,
Hennie.Brugman@mpi.nl

Résumé. La relation *voir/employé pour* d'un thesaurus est souvent plus complexe que la (para-)synonymie recommandée par l'ISO-2788, standard décrivant le contenu de ces vocabulaires contrôlés. Le fait qu'un non descripteur puisse renvoyer à plusieurs descripteurs (seuls les descripteurs sont pertinents dans le cadre de l'indexation contrôlée) fait que cette relation est complexe à utiliser dans un contexte d'annotation automatique : elle génère des cas d'ambiguité. Dans ce papier, nous présentons CARROT, un algorithme que nous avons mis au point pour classer les résultats de notre chaîne de traitements pour l'Extraction d'Information, et son utilisation dans le cadre de la sélection du descripteur pertinent lorsque plusieurs choix sont possibles. Cette sélection s'adresse à des documentalistes, dans le but de simplifier et d'accélérer leur travail, et se base sur la structure de leur thesaurus. Nous arrivons à un succès de 95 % dans nos suggestions ; nous discutons ces résultats et présentons des perspectives à cette expérimentation.

Abstract. The *use/use for* relationship a thesaurus is usually more complex than the (para-) synonymy recommended in the ISO-2788 standard describing the content of these controlled vocabularies. The fact that a non preferred term can refer to multiple preferred terms (only the latter are relevant in controlled indexing) makes this relationship difficult to use in automatic annotation applications : it generates ambiguity cases. In this paper, we present the CARROT algorithm, meant to rank the output of our Information Extraction pipeline, and how this algorithm can be used to select the relevant preferred term out of different possibilities. This selection is meant to provide suggestions of keywords to human annotators, in order to ease and speed up their daily process and is based on the structure of their thesaurus. We achieve a 95 % success, and discuss these results along with perspectives for this experiment.

Mots-clés : désambiguisation sémantique, algorithme de classement, annotation automatique.

Keywords: word sense disambiguation, ranking algorithm, automatic annotation.

1 Introduction

Thesauri are controlled vocabularies, often used for indexing and retrieving documents from collections. The standard thesauri contain two types of elements, preferred and non preferred terms, related with a link called *use/use for*. This link is considered as (para-)synonymy in the ISO-2788 standard (ISO, 1986) and can thus be useful for (semi-) automatic indexing applications : it enables a program to index a document with a preferred term (which is the type of thesaurus based controlled annotation we are interested in) either if the document contains an occurrence of the preferred term or if it contains occurrences of the corresponding non preferred term. In reality, this *use/use for* relationship is often more complex, and can generate ambiguity problems when used "as is" in an automatic application. We present in this paper the solution that we have developed in our project for selecting the relevant preferred term, given an occurrence of an ambiguous non preferred term in a text. This selection algorithm is based on the thesaurus's structure. The thesaurus we used in this experiment is the GTAA, which is employed for indexing and retrieving TV programs at the Netherlands Institute for Sound and Vision, the Dutch national TV archives. Our project, CHOICE[1], is collaborating with this Institute and focuses on easing and speeding up the work of cataloguers by providing them with a ranked set of keywords referring to their thesaurus' entries as indexing suggestions. We will present our project's goal and the specificity of this use case in the following section (section 2), followed by a description of thesauri in general and the GTAA itself (section 3). In this section, we will show the different semantics of the *use/use for* relationships and the problem of having multiple links between preferred and non preferred terms. We then present our annotation pipeline (section 3.4), including the algorithm that we elaborated to rank the extracted keywords, and that we propose here for selecting the relevant preferred term out of multiple possibilities (section 3.5). Section 5 shows our experiment to evaluate this algorithm in this Word Sense Disambiguation context. We achieved a 95 % of success, but are still facing minor and more important problems. We discuss them and conclude with perspectives for this experiment in section 6.

2 The CHOICE project

Charting the Information Landscape Employing Context Information, the CHOICE project deals with the suggestion of metadata from textual resources to annotate video documents. In the context of the Dutch TV archives, the cataloguers check a set of textual documents, on top of watching the program itself, to make their descriptions. One of the goals of our project is to build on existing Information Extraction platforms, extend and tune them to our specific needs in order to cope with the particularities of this specific use case and provide the cataloguers with a relevant set of keywords as indexing suggestions. Our Information Extraction is based on the content of the thesaurus that they are currently using at Sound and Vision, enriched and transformed by us. We present this thesaurus in the following section, and the specificity of our task in the section describing our ranking algorithm.

[1] http://www.nwo.nl/CATCH/CHOICE

3 The GTAA thesaurus

3.1 A thesaurus according to the ISO 2788 standard

A thesaurus is *The vocabulary of a controlled indexing language, formally organized so that the* a priori *relationships between concepts (for example as "broader" and "narrower") are made explicit.*[2]

Although this definition mentions *concepts*, a thesaurus contains terms (preferred and non preferred terms), organized according to 5 relationships : broader term (BT), narrower term (NT), related term (RT), use (US[3]) and use for (UF). A preferred term is *A term used consistently when indexing to represent a given concept* [...][4], whereas a non preferred term should not be used for indexing, but is only useful at search time to point different words possibly expressing the same idea towards the one that has been chosen to represent it in the thesaurus. In practice, as we detail in the following section, this relationship can encode different kinds of links, as suggested by these examples : hurricane UF cyclone, insurgent UF guerilla's, organ UF church organ, oven UF magnetrons, octopus UF calamary.

The other relationships should stand only between preferred terms. BT relates a term with a more generic one, supposed to index a larger set of documents. For example *Means of transportation* is a BT of *Bus*. NT is the relationship between a term and a more specific one, that should be used to index a subset of the documents indexed by the more generic one (*Bus* and *School bus*). RT is a non hierarchical relationship between two terms in the same domain, as *Bus* and *Driver*, for example.

3.2 The GTAA

The GTAA thesaurus, a Dutch acronym for "Common Thesaurus for Audiovisual Archives", is the controlled vocabulary used for the Sound and Vision documentation process. It contains approximately 160.000 terms. They are divided in 6 disjoint facets : Keywords (about 3800 preferred terms), Locations (about 14.000), Person Names (about 97.000), Organization-Group-Band Names (about 27.000), Maker Names (about 18.000) and Genres (113 preferred terms). The thesaurus mainly uses constructs as presented in the ISO 2788 standard and commonly used in companies or institutions : amongst others, use, use for, broader term, narrower term, related term. Terms from all facets of the GTAA may have related terms and use for relationships, but only Keywords and Genres can also have broader term/narrower term relations, organizing them into a set of hierarchies. Additionally, Keyword terms are thematically classified in 88 subcategories of 16 top Categories (Nature, Society,...). Although the data model that is used for the thesaurus allows links between terms across facets, no instances of these links currently exist. This experiment concerns only automatic indexing with terms from the Keyword facet.

[2](ISO, 1986), section 3-Definitions.
[3]In general this relationship is encoded USE, but this acronym is the one used in the GTAA.
[4](ISO, 1986), section 3-Definitions.

3.3 Different semantics and non uniqueness of the Use relationship

In the GTAA, there are 1377 US relationships, *i.e.* 1377 times a non preferred term is associated with a preferred term. Some of these non preferred terms are associated with multiple different preferred terms and some of the preferred term are associated with multiple non preferred terms. In the first case, the non preferred terms is polysemic or has different domains of application, each meaning or domain having an explicit preferred term : for example, the non-preferred term minority[5] has two preferred terms, ethnical minority and religious minority. In the latter case (one preferred term associated with different non preferred terms), different notions were grouped under one common and single preferred term. This is either done for easing the thesaurus' use (the fewer terms there are, the easier it is to find the most appropriate one when indexing), or because the distinction was not relevant for indexing the TV programs of Sound and Vision : for example, the preferred term diplomats groups two non preferred terms, ambassadors and consuls. When having a close look at the nature of the US, UF relationship we see four different types :

– Synonyms : To cleanse US To clean
– Meronym : Sabbath US Jewish religion
– Hyponym : Scanner US Hardware
– Semantically related : Geiger counter US radioactivity

83 non preferred terms are associated with more than one preferred term in the thesaurus, ranging from 2 to 3 different preferred terms. This non unique association can be a source of problems when using the thesaurus' content as a basis for automatic indexing. If we select the wrong preferred term, we might for example suggest petrol (aardolie) as an indexing term for a document about food, because the non preferred term oil (oliën) has both petrol and vegetable oil (plantaardige oliën) as its preferred term. We will present in the next section our semi-automatic annotation pipeline, the ranking algorithm applied to the term extraction, CARROT, and its usefulness for selecting the right preferred term out of 2 to 3 different possibilities.

3.4 Semi-automatic annotation pipeline

3.4.1 The pipeline

As stated in section 2, the goal of the semi automatic annotation pipeline is to suggest appropriate indexing terms to cataloguers, with the goal of easing their job and increasing their productivity. From discussion with the cataloguers it followed that they like a focussed and limited set of keywords : focussed because they only experience a suggestion as supportive if it closely matches the main topic of the document, limited because actual work process of cataloguers only allows for a limited number of terms to be attached to a document. Another reason for that requirement is that the inspection of the suggested terms should improve the work process, so the inspection time and the mental processing of the suggestions need to be bounded in order not to generate additional burdens.

The pipeline consists of tree parts : a term detector, a term collector and a term ranker. As input to our pipeline we use our selected corpus and the GTAA. The output of the pipeline is a ranked

[5] All the terms we mention in this paper are translated from Dutch to English out of consideration for our readers. We tried to select examples which have the same ambiguity in their semantics in the English translation

list of GTAA preferred terms.

3.4.2 The input : GTAA in a RDF-OWL representation

As input we use an RDF-OWL representation of the GTAA, based on the SKOS Working Draft (see (van Assem *et al.*, 2006) and (Miles & Brickley, 2005)). The SKOS representation of a thesaurus is "concept based" : instead of terms, the entities are nodes with identifiers (ID), to which labels are attached, a prefLabel to represent the preferred term, and one or more altLabel(s) to represent the non preferred term(s). As the GTAA entries are in plural form, we also extended this model to add the information of the singular form corresponding to the original thesaurus terms. This model has drawbacks, and has an obvious conceptual bias, but it helps gathering pragmatically different strings corresponding to the same annotation ID. These strings are called "textual representations of the concept" in the GATE pipeline, and we decided to keep this terminology here.

3.4.3 The term detector : GATE with the Apolda plug-in

The term detector scans a text and looks for all possible textual representations of concepts. The detector is built with the Apolda plug-in in GATE architecture (Maynard *et al.*, 2003). After tokenization, the Apolda plug-in makes a simple string matching. It annotates a piece of text with the ID of the "concept" corresponding to the longest matching textual representation. If for a piece of text multiple concepts have the same longest matching textual representation, which can be the case for a non preferred term with multiple preferred terms, the plug in generates all possible annotations. This means that the string minority will receive two annotations : Keyword_ethnical_minority and Keyword_religious_minority. The string religious minority however will only receive the latter. The term detector is not case sensitive.

3.4.4 The term collector

The outcome of the term detector is an annotated text. In this text, multiple annotations can correspond to the same "concept". The term collector collects all the annotation ID's, computes their number of occurrences and writes the output into one file.

3.5 The term ranker and WSD algorithm : CARROT

The file with ID's and number of occurrences computed at the previous step is fed into the Cluster And Rank Related to Ontology and Thesauri algorithm (CARROT algorithm) (Gazendam *et al.*, 2006).

CARROT uses the fact that terms in the Keyword facet of the GTAA are related to others via the related term, broader term and narrower term relations. We hypothesise that terms which relate to a lot of the other terms found in the text can be semantically more representative of the core topics of the TV program than terms which are found more often but without any relations to others. If one of the thesaurus relationships exists between two of the found terms we say that a relation of distance 1 exists. We also check if an intermediate term connects two terms in the

GTAA. These connections via intermediate terms are defined as relations of distance 2. We do not make any distinction in the type of relationships.

To rank the extracted keywords, we use the following rules :
– Step 1. We select the keywords with both a distance 1 and a distance 2 relation. We then order these keywords based on their number of occurrences, putting the most frequent on top of the list.
– Step 2. We select the remaining keywords with a distance 2 relation to keywords found during Step 1. We order these keywords based on their number of occurrences and add them to the list.
– Step 3. We select the remaining keywords with a relation. We order these keywords based on their number of occurrences and add them to the list.
– Step 4. We order the remaining keywords based on their number of occurrences and add them to the list.

This algorithm creates clusters of ranked terms (several terms can have the same rank, they are then simply ordered alphabetically).

Our previous experiment in (Gazendam *et al.*, 2006) showed that only the top clusters provided relevant keywords, so we intend to present the cataloguers with only these top clusters by default, with the possibility to access the whole ranked list if they wish to. In this paper we propose this CARROT algorithm as a means for selecting the right preferred term (right interpretation) for a non preferred term with multiple preferred terms (an ambiguous word). For example, the text :" *Snacks do not contain a lot of minerals.*" contains the non preferred term minerals and the preferred term snacks. minerals has three preferred terms : food, fertilizer and ore. All are considered to occur once, because their common non preferred term occurs once. These three plus snacks are fed into CARROT. Due to the direct relation between the terms food and snacks, food now ranks higher than the other two preferred terms. This means that we here interpret minerals as referring to food in this case.

As the non preferred term attributes the same number of occurrences to all its preferred terms, three scenarios are possible :
– One of the preferred terms has more direct or indirect relations to other found terms and ranks higher as a result ;
– One of the preferred terms combines a higher number of occurrences due to the fact that the preferred term appeared itself in the text or one of its other non preferred terms appeared in the text ;
– The different preferred terms rank equally high.

The output of the pipeline is the same list of annotation ID's as the input, but ranked. Therefore, our hypothesis for Word Sense Disambiguation is that the irrelevant preferred terms will not be connected to any of the other found keywords, and thus will be ranked at the bottom of the list. As a consequence, they will not be shown to the cataloguers as indexing suggestion. We present the positioning of our experiment with the state of the art in Word Sense Disambiguation in the following section, followed by the experiment itself.

4 Related Work

The task we are interested in in this paper can be related to Word Sense Disambiguation. In (Ide & Véronis, 1998), the authors describe the typical two-step process for this task :

1. Define the set of senses per lexical unit ;
2. Use either a context-based method to determine which of the senses corresponds to the occurrence of the lexical unit considered, or an external knowledge source.

Many works mention the use of a dictionary as an external knowledge for that purpose ((Veronis & Ide, 1990), for example), whereas statistically-based or machine-learning methods advertise the corpus-based contextual approach (see for example (Yarowsky, 1995)). Of course, some mixed approaches exist, as (Stevenson & Wilks, 2001). In our use case, the set of senses to take into account is the set of possible preferred terms for each ambiguous non preferred term. The method that we experiment here is using external knowledge, but instead of the lexical content of dictionary definitions, or instead of trying to map the lexical environment of the external knowledge to the corpus content, we use the thesaurus independently, and take only into account the number of occurrences of each term as a contextual information. The selection of the relevant sense, *i.e.* of the relevant preferred term, is made only based on relationships crafted by hand by cataloguing experts when building the thesaurus. Therefore it is still different from (Yarowsky, 1992), who also based his Word Sense Disambiguation algorithm on a thesaurus.

5 The experiment

5.1 Experiment : selecting the right keyword when multiple USE relations are possible

For this experiment, we annotated our documents with all the possible preferred terms related to the non preferred terms we found in the texts, along with their number of occurrences, and we will check whether the algorithm designed for ranking the IE output will help us disambiguating between the different possibilities. We will evaluate whether CARROT

1. Ranks the relevant preferred term higher ;
2. Ranks the irrelevant preferred terms low enough for them not to be part of the keywords suggested to the cataloguers.

5.2 Material

We constructed our corpus from a set of over 500 catalogue descriptions from Sound and Vision, related to TV programs. Each of these catalogue descriptions contains specific fields, that are described in Dublin Core : *e.g.* maker, title and keywords. One of the fields is a free text description called summary. In the Keyword field the topic of the program is described by a limited set of preferred terms from GTAA's Keyword facet. From this set of catalogue descriptions we selected all files which :

1. contain a non preferred terms which has multiple preferred terms and
2. have one of its related preferred terms appear in the keyword field

Based on these requirements we selected automatically a corpus of 121 documents, of averagely 200 words each. The second requirement is related to evaluation purpose : the preferred term that was chosen to describe the document can be seen as the correct interpretation of the non preferred term present in the description text. We base ourselves on this assumption to evaluate

the results of our ranking algorithm : the preferred term present in the Keyword field should be ranked higher than the other possible preferred terms.

5.3 Experiment

We ran our pipeline on our corpus. After completion we looked at the non preferred term, the rank of all associated preferred terms in the ranked list and compared this ranked list with the preferred term in the Keyword field of the catalogue description. We have three possible outcomes of this comparison :

1. Correct suggestion : the suggested preferred term[6] is the preferred term in the keywords
2. Wrong suggestion : the suggested preferred term is not the preferred term in the keywords
3. Undecidable : No suggestion is made because two (or all three) preferred terms rank equally high

When evaluating the results, we also came across a set of unusable data. We discuss this point in the following section. The results are shown in table 1

correct	undecidable	wrong	unusable data	total
43	26	2	50	121

TAB. 1 – Results

5.3.1 Discussion

One of the issues that arose when evaluating our results was that we still have numerous unusable documents in our corpus : it turned out that for some documents, the non preferred term is found in the keyword field. According to the production rules of Sound and Vision a non preferred term cannot be used in the keyword field, but the set of keywords changes over time : a preferred term may be ambiguous and as a consequence be changed to a non preferred term. Because we used old descriptions in our corpus, some of these contained previously preferred terms which now became non preferred ones in their keyword field. This is the case, for example, for murder assault (8 occurences) and tent kampen (23 occurences). These two examples account for two thirds of the unusable data. We excluded these from our analysis.

For the remainder of the corpus, in approximately 19 out of 20 (95%) cases, the suggestions are not incorrect. We found only two cases in which we gave a wrong suggestion. Both mistakes are with the same non preferred term clubs which has as preferred terms hotel, restaurant and cafe (HRC) and association. This word club was used in the context of football clubs. One text was on the share issue of soccer club Ajax. The other text was on the showing of a documentary on the soccer club Ajax in a theater. The term club had the meaning of association in both cases, referring to the soccer association. However the hotel, restaurant and cafe was suggested. In both cases terms at distance 2 from HRC were present in the text : theater via the intermediate term nightlife and director via the intermediate term enterprice. On the other hand associations did not have direct or distance 2 connections to other extracted terms in the football domain as soccer, supporter, match, trainer : the distance in the thesaurus between

[6] *i.e.* the preferred term with the highest rank in the list.

these terms and association was too big. In our corpus, we have two other instances of club for which the matching to its preferred terms is successful once and undecidable another time. Both these texts where also in the soccer domain and having the preferred term association.

This could suggest that we have one "preferential preferred term" in the corpus, and that this information could be used for solving in a light way the ambiguity problem. Unfortunately, this is not always the case : the non preferred term windmills occurs once as wind turbine and once as mill ; in both cases the correct suggestion is made by our system. Other non preferred terms with a bigger number of occurrences also have a non regular distribution of their preferred terms.

Another remarkable feature of the results is the big number of undecidable cases. The reason why we encounter this big number of undecidable cases is manyfold :

1. Our method uses general conditional rules. These conditions are not really specific : *having any distance 1 relation satisfies a condition*. As a result, in many cases both preferred terms fit the same conditions. This can be amended by sharpening these conditions, for example by counting the number of terms at distance 1 or distance 2.
2. The texts of our corpus are relatively small, so the number of found (and related) terms is also small, and the number of occurrences too low to disambiguate between different possibilities.
3. In many cases the different preferred terms have a distance 1 relation to other extracted terms, increasing the chance of a tie. At the same time this means that the difference in meaning between the preferred terms can be subtle, giving value to the undecidability. For example, it is very difficult and maybe not relevant to distinguish between the three preferred terms related to toxin, namely poison, venom and dangerous substance, in the context of a TV program about farmers getting ill after using a toxin as a form of herbicide.

The last remark that we can make is that, due to the small number of different keywords in the different texts, very few clusters were created. As a consequence, it was hardly ever the case that the non relevant preferred terms found place low enough in the ranked list not to be proposed for indexing suggestion. Therefore, we should modify our algorithm in order to make it take into account only the preferred term with higher rank, and remove the other related preferred terms from the suggestion list.

6 Conclusion and Perspectives

We investigated whether our method and the CARROT algorithm could be used for disambiguation in an indexing setting. In cases of ambiguity, it only gives suggestions for which preferred term to choose in two cases out of three, but when it gives a suggestion, it is correct so in approximately 19 out of 20 cases. The two bad suggestions came from the same thesaurus concept, and were due to its lack of structure. Using another external resource like the Princeton University's WordNet thesaurus could help us cope with that problem. However, the interpretation of our success rate and percentage of undecidable cases must be subject of study : it is up to the cataloguers to determine whether these numbers are fair [7]. This is the subject of another study, that we will also conduct in the course of our project.

[7] A success of 19 out of 20 seems quite reasonable in the perspective of IR publications, but when talking about automatically securing railway crossings, the same success ratio is considered really bad.

Acknowledgements

This research was partly supported by the NWO funded CATCH program, including the CHOICE project. We want to thank our colleagues, both at the University and at Sound and Vision for their daily help, support and our fruitful collaboration.

Références

GAZENDAM L., MALAISÉ V., SCHREIBER G. & BRUGMAN H. (2006). Deriving semantic annotations of an audiovisual program from contextual texts. In *Proceedings of First International workshop on Semantic Web Annotations for Multimedia (SWAMM 2006)*.

IDE N. & VÉRONIS J. (1998). Introduction to the special issue on word sense disambiguation : The state of the art. *Computational Linguistics*, **24**(1), 1–40.

ISO (1986). *Documentation - guidelines for the establishment and development of monolingual thesauri*. International Organization for Standardization, ISO 2788-1986 edition.

MAYNARD D., TABLAN V. & CUNNINGHAM H. (2003). Ne recognition without training data on a language you don't speak. *ACL Workshop on Multilingual and Mixed-language Named Entity Recognition : Combining Statistical and Symbolic Models*.

MILES A. & BRICKLEY D. (2005). Skos core guide. 2nd W3C Public Working Draft.

STEVENSON M. & WILKS Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, **27**(3), 321–349.

VAN ASSEM M., MALAISÉ V., MILES A. & SCHREIBER G. (2006). A method to convert thesauri to skos. In *Proceedings of the Third European Semantic Web Conference (ESWC'06)*.

VERONIS J. & IDE N. M. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th conference on Computational linguistics*, p. 389–394, Morristown, NJ, USA : Association for Computational Linguistics.

YAROWSKY D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of 14th International Conference on Computational Linguistics (COLING-92)*.

YAROWSKY D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics (ACL' 95)*.

Repérage de sens et désambiguïsation dans un contexte bilingue

Marianna APIDIANAKI
Lattice, Université Paris 7, CNRS
ENS-1 rue Maurice Arnoux, F-92120, Montrouge
Marianna.Apidianaki@linguist.jussieu.fr

Résumé. Les besoins de désambiguïsation varient dans les différentes applications du Traitement Automatique des Langues (TAL). Dans cet article, nous proposons une méthode de désambiguïsation lexicale opératoire dans un contexte bilingue et, par conséquent, adéquate pour la désambiguïsation au sein d'applications relatives à la traduction. Il s'agit d'une méthode contextuelle, qui combine des informations de cooccurrence avec des informations traductionnelles venant d'un bitexte. L'objectif est l'établissement de correspondances de traduction au niveau sémantique entre les mots de deux langues. Cette méthode étend les conséquences de l'hypothèse contextuelle du sens dans un contexte bilingue, tout en admettant l'existence d'une relation de similarité sémantique entre les mots de deux langues en relation de traduction. La modélisation de ces correspondances de granularité fine permet la désambiguïsation lexicale de nouvelles occurrences des mots polysémiques de la langue source ainsi que la prédiction de la traduction la plus adéquate pour ces occurrences.

Abstract. Word Sense Disambiguation (WSD) needs vary greatly in different Natural Language Processing (NLP) applications. In this article, we propose a WSD method which operates in a bilingual context and is, thus, adequate for disambiguation in applications relative to translation. It is a contextual method which combines cooccurrence information with translation information found in a bitext. The goal is the establishment of translation correspondences at the sense level between the lexical items of two languages. This method extends the consequences of the contextual hypothesis in a bilingual framework assuming, at the same time, the existence of a semantic similarity relation between words of two languages being in a translation relation. The modelling of fine-grained correspondences allows for the disambiguation of new occurrences of the polysemous source language lexical items as well as for the prediction of the most adequate translation for those occurrences.

Mots clés : désambiguïsation contextuelle, similarité sémantique, substituabilité, traduction.

Keywords: contextual disambiguation, semantic similarity, substitutability, translation.

1. Désambiguïsation lexicale pour la traduction

La définition de la nature des sens, leur énumération et leur description constituent des questions centrales dans la problématique de la désambiguïsation lexicale, auxquelles une

réponse unanime est loin d'être trouvée. Les besoins concernant le degré de désambiguïsation ainsi que le type et le niveau des distinctions sémantiques varient dans le cadre de différentes applications du Traitement Automatique des Langues (TAL). Ainsi, les informations trouvées dans des ressources sémantiques prédéfinies s'avèrent souvent peu conformes aux besoins des applications particulières, et les méthodes de désambiguïsation sont parfois critiquées pour ne pas être liées à une application réelle.

Dans cet article, nous allons présenter une méthode originale de désambiguïsation lexicale qui opère dans un contexte bilingue et dont les résultats sont, par conséquent, utilisables dans des applications relatives à la traduction. Il s'agit d'une méthode de cooccurrences qui peut opérer sur les deux côtés d'un corpus parallèle (bitexte) : les contextes de la langue source (LS) et les contextes de la langue cible (LC). La combinaison d'informations contextuelles et traductionnelles permet le repérage de distinctions sémantiques au sein des mots polysémiques et l'établissement, entre les mots des deux langues, de correspondances au niveau sémantique exploitables dans des systèmes de traduction automatique ou assistée par ordinateur.

2. Principes de la méthode et hypothèses sous-jacentes

Les hypothèses théoriques sous-jacentes à cette méthode de désambiguïsation lexicale sont les suivantes :

1. l'hypothèse contextuelle du sens (Firth, 1957 ; Harris, 1985), d'après laquelle le sens des mots correspond à leurs usages dans les textes ;
2. l'hypothèse de l'existence d'une relation de similarité sémantique entre les mots de deux langues entretenant une relation de traduction dans des textes réels.

D'après la première hypothèse, l'analyse du contexte lexical (co-texte) entourant un mot dans des textes peut éclairer sa sémantique. Le contexte lexical a été exploité pour la désambiguïsation aussi bien dans des méthodes qui procèdent à la sélection du bon sens des mots à partir d'un dictionnaire dans un cadre monolingue (Lesk, 1986) et bilingue (Brun *et al*., 2001 ; Dufour, 1997), que dans des méthodes de désambiguïsation n'utilisant pas de ressources lexicales préalables. De telles méthodes, proposées dans un cadre monolingue, sont celles de Schütze (1998), de Véronis (2003) et de Pantel et Lin (2002) ; les deux premières exploitent les informations de cooccurrence des mots, tandis que la troisième met l'accent sur le contexte syntaxique. Dans un cadre de traduction, le contexte lexical des mots est exploré dans les méthodes de désambiguïsation proposées par Brown *et al*. (1991)[1], Kaji *et al*. (2003)[2] et Specia *et al*. (2006)[3], tandis que celle proposée par Dagan et Itai (1991) exploite le contexte syntaxique de la LC pour choisir le bon équivalent de traduction.

La deuxième hypothèse, citée plus haut, postule que, dans le cas de correspondances au niveau lexical au sein d'un corpus parallèle, le sens véhiculé par un équivalent de traduction est supposé similaire à celui du mot source qu'il traduit. Par conséquent, les équivalents de traduction possibles des mots polysémiques de la LS sont censés traduire les différents sens

[1] La méthode utilise des questions binaires pour choisir entre deux sens d'un mot.
[2] La méthode exploite des corpus comparables et utilise des informations sur l'alignement translinguistique de paires de mots liés. Les sens sont décrits par un ou plusieurs équivalents, dont le regroupement se base sur la similarité distributionnelle dans les deux langues. Cette méthode ne prend pas en compte les ambiguïtés parallèles entre les mots des deux langues, et elle présuppose que chaque équivalent traduit uniquement un sens du mot polysémique.
[3] La méthode proposée par Specia *et al*. prend en compte des informations de cooccurrence lexicale dans la LC acquises à l'aide de requêtes effectuées sur le Web concernant des fragments de texte de la LC.

de ces mots dans la LC, sens reflétés aussi dans le contexte lexical de la LS. La nouveauté de notre approche consiste justement au repérage automatique des sens des mots polysémiques par projection des informations de cooccurrence d'un côté du bitexte à l'autre, sans recours à une ressource lexicale préalable. L'analyse sémantique qui en résulte concerne tant les mots de la LS que les équivalents, et les résultats sont directement exploitables pour la désambiguïsation et la sélection lexicale dans la traduction.

La correspondance sémantique entre deux unités lexicales en relation de traduction peut être mise en évidence et servir à la modélisation de correspondances sémantiques au sein d'un système automatique. La correspondance à laquelle nous faisons face avant la désambiguïsation d'un mot polysémique de la LS se situe au niveau lexical, où le mot en question correspond à plusieurs équivalents dans la LC. Le but est de raffiner cette relation et de représenter les liens entre les mots des deux langues à un niveau d'analyse plus élevé. Sur la base des hypothèses précitées, nous acceptons que les informations venant du co-texte des occurrences de l'unité source qui sont traduites par un équivalent précis dans le corpus éclairent tant le(s) sens véhiculé(s) par ces occurrences que celui(ceux) de l'équivalent de traduction. Le co-texte du mot source par rapport à un équivalent précis correspond aux mots de contenu (noms, adjectifs et verbes) qui cooccurrent avec le mot dans les segments de traduction où il est traduit par cet équivalent[4].

Dans une méthode contextuelle monolingue de désambiguïsation, la comparaison des contextes (lexicaux ou grammaticaux) des occurrences du mot polysémique permet leur clusterisation en fonction de leur similarité et les clusters résultants sont censés illustrer les différents sens du mot. Ici, au lieu de comparer entre eux et de clusteriser les contextes dans lesquels un mot polysémique apparaît, nous allons comparer entre eux et clusteriser des ensembles de contextes correspondant à chacun de ses équivalents. Dans le paragraphe suivant, nous allons décrire comment les ensembles en question sont construits.

3. Prétraitement du corpus

Le corpus utilisé dans ce travail pour l'apprentissage est un bitexte anglais-grec de 4 000 000 mots (Gavrilidou *et al.,* 2004), lemmatisé, morphosyntaxiquement étiqueté et aligné au niveau des phrases et au niveau des mots (Simard, Langlais, 2003). La source principale des textes est le *Journal de l'Union Européenne* (domaines : droit [42 % des textes du corpus], santé [24 %], éducation [21 %]), mais il y a aussi des textes venant de l'Office National Hellénique du Tourisme (11 %), ainsi qu'un petit nombre de textes scientifiques sur l'environnement (2 %). Pour chaque mot polysémique étudié, un sous-corpus a été créé, qui contient les segments de traduction dans lesquels le mot source occurre[5]. Le choix des segments de traduction en tant que contexte est dicté par notre objectif d'exploration de l'influence du co-texte proche des unités source sur la désambiguïsation et le transfert lexical. Les segments constituant le sous-corpus d'un mot polysémique ont été regroupés en fonction de ses équivalents de traduction. Ainsi des ensembles de phrases correspondant à chacun des équivalents sont créés dans les deux côtés du bitexte, comme cela est décrit dans la figure 1.

Cette figure illustre le sous-corpus d'un mot source *m*, qui est traduit dans le corpus par trois équivalents différents : *a*, *b* et *c*. Dans la partie gauche de la figure, nous avons les phrases de la LS et, à droite, leurs traductions dans la LC, qui se trouvent dans les mêmes segments de

[4] Tous les calculs opèrent sur les lemmes (*types*) auxquels les mots des contextes (*tokens*) ont été ramenés.
[5] Un segment peut contenir de 0 à 2 phrases par langue. Par exemple, un alignement *2:1* met en correspondance 2 phrases du texte de la LS avec 1 phrase du texte de la LC, à l'intérieur d'un segment.

traduction, comme cela a été déterminé par le processus d'alignement des phrases.

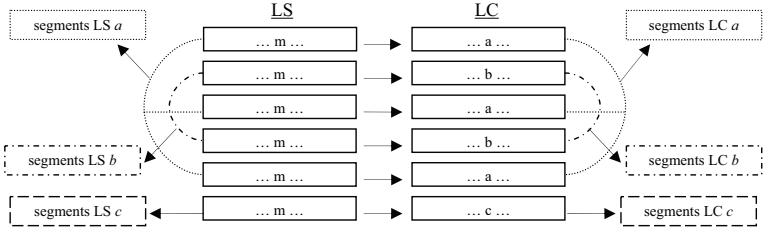


Figure 1 : Regroupement des segments de traduction en fonction des équivalents

Les segments sont regroupés en fonction des équivalents. Ainsi, nous avons un ensemble de phrases dans la LS correspondant aux occurrences de *m* traduites par *a* et un ensemble de phrases dans la LC correspondant à ces traductions et contenant, bien évidemment, le mot *a*. Nous procédons de la même manière pour les autres équivalents (*b* et *c*) en constituant les groupes de phrases respectifs dans les deux langues[6]. Ces ensembles de segments constituent l'entrée de la méthode de similarité sémantique qui sera présentée par la suite.

4. Méthode d'estimation de similarité sémantique

4.1 Description et présupposés

Le calcul de similarité sémantique est appliqué sur les ensembles de segments correspondant aux équivalents de traduction d'un mot polysémique. Il peut porter sur les contextes du mot source aussi bien que sur les contextes des équivalents. Sur la base de l'hypothèse contextuelle du sens, un grand degré de similarité des contextes de la LS indique l'homogénéité sémantique du mot source, tandis que leur dissimilarité constitue un indice de l'existence de distinctions sémantiques au sein du mot en question. D'après la deuxième hypothèse du départ, qui concerne l'existence d'une correspondance d'ordre sémantique entre les mots de deux langues entretenant une relation de traduction, la similarité plus ou moins grande des contextes de la LS correspondant à des équivalents différents indique le degré de similarité sémantique des équivalents en question. Ainsi, les résultats du calcul de similarité au sein de la LS permettent la clusterisation (ou la distinction) des équivalents en fonction de leur similarité (ou dissimilarité) sémantique.

Les équivalents clusterisés sur la base de leur similarité sémantique sont censés traduire le même sens du mot polysémique[7]. Dans une approche contextuelle de la similarité sémantique, les mots similaires sont considérés comme substituables au sein des contextes qui induisent leur relation (Miller et Charles, 1991). Ainsi, il est possible d'émettre l'hypothèse que lorsque le calcul porte sur les contextes de la LS, les équivalents clusterisés sont substituables en tant que traductions pour les occurrences du mot source trouvées dans les contextes révélant leur similarité.

La projection des résultats de la clusterisation sur le mot source permet donc le repérage des sens véhiculés par le mot. Ainsi la notion de similarité sémantique basée sur la similarité contextuelle, et hautement utilisée dans le cadre de la désambiguïsation monolingue, est

[6] Dans la figure, nous décrivons ces ensembles de phrases comme « segments LS *équivalent* » et « segments LC *équivalent* ».
[7] En revanche, la distinction des équivalents signale leur dissimilarité sémantique, c'est-à-dire qu'ils traduisent des sens différents du mot source.

reprise ici afin d'émettre des jugements sur la similarité des équivalents de traduction qui peuvent, par la suite, être projetés sur les unités polysémiques source. Un avantage de l'utilisation de cette méthode par rapport à l'application d'une méthode monolingue de désambiguïsation (Apidianaki, 2006) est que les distinctions proposées sont beaucoup plus pertinentes pour la traduction[8].

4.2 La mesure de similarité sémantique

La mesure utilisée pour estimer la similarité sémantique entre les équivalents de traduction est une variation de la mesure de Jaccard pondérée (JP), proposée par Grefenstette pour la création de thesaurus sémantiques dans un contexte monolingue (1994 : 48-50)[9]. Dans notre travail, les informations utilisées pour le calcul concernent la cooccurrence des mots à l'intérieur des segments de traduction. Le contexte de la LS par rapport à un équivalent est constitué par les mots de contenu cooccurrant avec le mot source dans les segments où il est traduit par cet équivalent précis. D'autre part, les traits sur lequel porte le calcul dans la LC sont les mots de contenu du co-texte de l'équivalent dans les traductions. Pour chaque trait nous calculons son **poids global** (global weight : *gw*) :

$$gw(trait_j) = 1 - \sum_i \frac{p_{ij} \log(p_{ij})}{nrels} \qquad \text{où} \quad p_{ij} = \frac{fréquence absolue du trait_j \; avec l'équiv_i}{nombre total de traits pour l'équiv_i}$$

$$\text{et} \quad nrels = le nombre total de relations extraites du corpus pour le trait_j$$

son **poids local** (local weight : *lw*) : $lw(équiv_i, trait_j) = \log(fréquence du trait_j \; avec l'équiv_i)$

et son **poids total** (whole weight : *W*) : $W = gw * lw$

Les éléments du contexte qui nous servent comme traits sont donc pondérés en fonction de leur dispersion dans les textes (*gw*) et de leur fréquence d'occurrence avec chaque équivalent précis (*lw*). Le poids global d'un trait est calculé en prenant en compte la somme de la probabilité d'occurrence du trait avec chacun des équivalents (p_{ij}), ainsi que le nombre total d'équivalents avec lesquels ce trait occurre (*nrels*). Le poids local est calculé sur la base de la fréquence d'apparition du trait avec un équivalent précis. Ainsi le coefficient pondéré nous permet d'attribuer une importance aux traits, qui est proportionnelle à leur pertinence pour l'estimation de la similarité entre les équivalents. Le Jaccard entre deux équivalents *m* et *n* est calculé par la formule suivante :

$$JP(équiv_m, équiv_n) = \frac{\sum_j \min(W(équiv_m, trait_j) W(équiv_n, trait_j))}{\sum_j \max(W(équiv_m, trait_j) W(équiv_n, trait_j))}$$

4.3 Analyse des résultats

Le processus a été appliqué sur dix mots polysémiques anglais[10]. Nous allons présenter ici un exemple illustrant le fonctionnement de la méthode et les résultats que nous pouvons obtenir. Cet exemple concerne le mot *movement* qui a neuf équivalents de traduction dans le corpus[11]: *κυκλοφορία* (251), *διακίνηση* (38), *κίνηση* (28), *μετακίνηση* (19), *κίνημα* (11), *κινητικότητα*

[8] Les sens proposés par l'application d'une méthode de désambiguïsation qui ne prend pas en compte les équivalents dès le début sont très nombreux, et il est difficile de créer des correspondances de traduction satisfaisantes au niveau sémantique. L'utilisation des équivalents comme indices pour le fusionnement des sens risque de nous faire entrer dans un cercle vicieux à cause de la polysémie propre aux équivalents mêmes.

[9] La similarité entre les mots est estimée sur la base de leurs contextes syntaxiques partagés.

[10] *Plant*, *movement*, *occupation*, *communication*, *treatment*, *passage*, *power*, *competence*, *facility*, *paper*.

[11] Entre parenthèses, nous donnons la fréquence avec laquelle l'équivalent traduit le mot source dans le corpus.

(6), *προσπάθεια* (1), *τάση* (1), *βήμα* (1). Les résultats du calcul de similarité entre ces équivalents sont décrits dans le tableau 1[12].

Paires d'équivalents		Contextes anglais	Contextes grecs
μετακίνηση	διακίνηση	**0,11**	**0,125**
κίνηση	διακίνηση	**0,099**	**0,13**
μετακίνηση	κίνηση	**0,087**	**0,141**
κυκλοφορία	διακίνηση	**0,078**	**0,091**
κίνηση	κυκλοφορία	**0,063**	**0,077**
κίνηση	κίνημα	**0,046**	0,052
	Moyenne	**0,043**	**0,062**

Tableau 1 : Relations de similarité entre les équivalents de *movement*

Les éléments des paires qui apparaissent au début de la liste sont censés entretenir des relations sémantiques plus fortes que ceux qui apparaissent vers la fin. L'analyse des résultats obtenus pour tous les mots étudiés a démontré que la moyenne des scores associés aux paires d'équivalents[13] pourrait constituer une sorte de seuil au-dessous duquel les relations trouvées ne sont pas pertinentes. Les relations repérées permettent la clusterisation des équivalents : les éléments inclus dans un cluster sont censés traduire le même sens du mot source tandis qu'un élément qui appartient à plusieurs clusters est supposé traduire des sens différents.

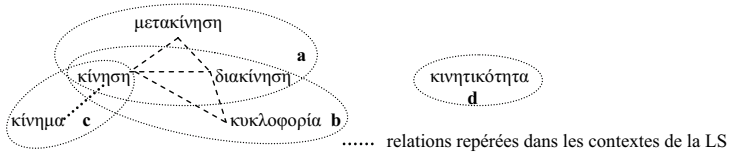


Figure 2 : Clusterisation des équivalents de *movement*

Les relations entre les équivalents sont induites soit par les contextes de la LS, soit par ceux de la LC, soit par les deux types de contexte ; ceci est illustré dans la figure 2 par l'utilisation de lignes différentes. D'après les résultats obtenus jusqu'à maintenant, les relations repérées seulement dans les contextes des traductions ne sont pas pertinentes, contrairement aux relations repérées dans les contextes des textes originaux ou dans les deux types de contextes.

La projection des clusters d'équivalents sur le mot polysémique source induit trois sens au sein du mot *movement*. Les deux premiers sens sont illustrés par les clusters *a* et *b* (*μετακίνηση-κίνηση-διακίνηση* et *διακίνηση-κίνηση-κυκλοφορία*) et correspondent aux usages de *movement* qui décrivent les notions de mobilité, de circulation et de transfert. Les sens décrits par ces deux clusters pourraient théoriquement être regroupés dans un sens plus large. Néanmoins, la distinction induite par les clusters séparés peut être expliquée comme reflétant des contraintes d'utilisation et de substitution des équivalents au sein des contextes. Le cluster *c* (*κίνημα-κίνηση*), décrit le sens métaphorique de *movement* (par ex. des mouvements sociaux et autres). Ce sens est le plus souvent traduit par *κίνημα*, mais aussi parfois par *κίνηση*. *Κίνηση* est un mot polysémique qui peut véhiculer, d'une part, le sens de mobilité et de déplacement physique et, d'autre part, le sens de mouvement social. L'isolation de certains équivalents, comme dans le cas de *κινητικότητα*, peut être due aussi bien à des raisons sémantiques qu'à leur basse fréquence d'occurrence dans le corpus, ce qui fait qu'il n'y a pas assez

[12] Paires faiblement liées (avec des scores inférieurs à la moyenne) : *μετακίνηση-κυκλοφορία* (0,035-0,05) / *-κινητικότητα* (0,039-0,009) / *-κίνημα* (0,009-0,044), *κινητικότητα-διακίνηση* (0,038-0,036) / *-κυκλοφορία* (0,022-0,023) / *-κίνημα* (0-0,045) / *-κίνηση* (0-0,029), *κίνημα-διακίνηση* (0,019-0,06) / *-κυκλοφορία* (0,008-0,018).

[13] La moyenne des scores attribués à toutes les paires (mêmes aux paires d'équivalents faiblement liés).

d'informations contextuelles pour les rapprocher des autres. Quand la fréquence d'occurrence d'un équivalent est grande par rapport à celle des autres équivalents et qu'il reste, malgré cela, isolé, nous pouvons supposer l'existence d'une différence sémantique. Dans le cas de *κινητικότητα*, son isolation est due à sa basse fréquence dans le corpus (6) et à son contexte très restreint, qui fait qu'il n'y a pas assez d'informations contextuelles sur cet équivalent qui pourraient le « rapprocher » des autres.

Les relations entre les équivalents contenus dans un cluster sont modélisées à l'aide des éléments contextuels qui les mettent en évidence, c'est-à-dire par les traits qui sont communs aux équivalents en question et que nous pourrions appeler leurs *co-textes assimilateurs* (Fuchs, 1994 :134-141). Ainsi chaque paire d'équivalents est caractérisée par l'ensemble de traits de leurs co-textes assimilateurs et par un ensemble de traits correspondant à chacun des équivalents, qui contient aussi bien leurs traits communs que les traits de leurs co-textes *dissimilateurs*, qui différencient l'un équivalent de l'autre. Il se peut que les co-textes dissimilateurs caractérisant un équivalent décrivent un sens véhiculé seulement par celui-ci (et non par l'autre membre de la paire). Le regroupement des traits dissimilateurs et assimilateurs au sein de ces ensembles provoque une perte d'informations relatives à ce type de distinctions qui n'a pas d'impact négatif important sur le processus de prédiction de traduction et qui améliore même les résultats du point de vue qualitatif. Le regroupement permet la sélection correcte de l'un des équivalents d'une paire, non pas seulement si celui-ci véhicule un sens bien distinct, mais aussi lorsqu'il est plus adéquat que l'autre équivalent de la paire pour traduire la nouvelle occurrence du mot source. En revanche, l'utilisation de correspondances de ce type dans le cadre d'une autre application, comme la recherche d'informations multilingues, nécessiterait probablement une modélisation plus fine des sens véhiculés par chacun des équivalents.

5. Évaluation

Le corpus de test utilisé pour l'évaluation de la méthode est différent du corpus d'apprentissage. Il s'agit de la partie anglais-grec du corpus parallèle EUROPARL (Koehn, 2003). Etant aligné au niveau des phrases, nous avons pu extraire à partir de ce corpus des segments de traduction contenant, d'une part, les phrases en anglais où les mots polysémiques occurrent et, d'autre part, leurs traductions en grec. L'évaluation de la méthode consiste à l'utilisation des correspondances établies entre les mots polysémiques et leurs équivalents de traduction pendant l'étape précédente pour (a) la désambiguïsation des nouvelles occurrences des mots polysémiques et (b) la prédiction des traductions les plus adéquates pour ces occurrences. Pour chacun des équivalents des mots polysémiques trouvés dans le corpus de test, nous avons retenu un ensemble de segments choisis au hasard. Par exemple, le mot *movement* est traduit dans EUROPARL par tous les équivalents clusterisés (voir figure 2) et nous avons retenu dix segments correspondant à chacun de ces équivalents.

Etant donné que l'apprentissage a opéré sur les formes des mots de catégories grammaticales précises et que les correspondances ont été modélisées à l'aide de celles-ci, il a fallu lemmatiser et étiqueter morphosyntaxiquement les nouveaux contextes pour rendre possible leur comparaison avec les informations retenues[14]. Voici un exemple d'un segment de traduction pour le mot *movement* :

[14] L'étiquetage morphosyntaxique et la lemmatisation ont été effectués à l'aide de l'étiqueteur TreeTagger (Schmid, 1994). Nous rappelons que seulement les noms, les adjectifs et les verbes des contextes sont retenus.

{I am therefore delighted that steps are finally being taken which will allow the theoretical freedom of movement of persons to be translated into practice, albeit still far from perfect practice! -- *Επομένως είμαι ικανοποιημένος που επιτέλους λαμβάνονται μέτρα τα οποία θα επιτρέψουν να γίνει πράξη η αρχή της ελεύθερης* ***κυκλοφορίας*** *των προσώπων, έστω και αν τα μέτρα αυτά είναι ακόμα, κατά τη γνώμη μου, ανεπαρκέστατα!*}

Le contexte de la nouvelle occurrence de *movement* est représenté de la manière suivante : {*be* (VBP), *delight* (VVN), *step* (NNS), *be* (VBG), *take* (VVN), *allow* (VV), *theoretical* (JJ), *freedom* (NN), *person* (NNS), *translate* (VVN), *practice* (NN), …}. La comparaison de cet ensemble d'éléments avec les traits qui caractérisent les correspondances entre *movement* et ses équivalents permet la sélection de l'équivalent le plus adéquat pour la nouvelle occurrence du mot. La prédiction de traduction avec le plus grand score pour cette occurrence de *movement* est la paire d'équivalents : *διακίνηση-κυκλοφορία* (score : 2,951).

La proposition de paires d'équivalents pour un mot constitue l'un des points forts de la méthode dans le sens où elle profite bien des informations paradigmatiques relatives à la similarité sémantique entre les équivalents, qui enrichissent les correspondances établies entre les mots des deux langues. La proposition d'une paire d'équivalents signifie qu'ils sont interchangeables au sein du nouveau contexte. Il faut souligner le rôle important des scores attribués aux nouveaux contextes par rapport aux correspondances établies ; ces scores sont calculés sur la base des poids attribués aux traits qui caractérisent les équivalents clusterisés. Ainsi, une proposition est faite non pas sur la base du nombre des traits communs entre le nouveau contexte et les correspondances modélisées, mais en fonction de la pertinence des traits du nouveau contexte par rapport aux éléments qui forment les correspondances.

Pour évaluer les résultats du point de vue quantitatif, nous utilisons les mesures de rappel et de précision ; le rappel correspond au rapport du nombre de prédictions correctes au nombre de traductions de référence et la précision au rapport du nombre de prédictions correctes au nombre de propositions faites par le système. Nous considérons comme prédiction correcte la proposition de l'équivalent qui traduit la nouvelle occurrence du mot dans le corpus de test – dans notre exemple il s'agit de l'équivalent *κυκλοφορία* – ainsi que les propositions d'équivalents qui retiennent une relation de similarité sémantique avec lui et donc se trouvent dans le même cluster[15]. La prise en compte des relations paradigmatiques établies entre les équivalents et décrites à l'aide des clusters apporte une amélioration des résultats de l'évaluation par rapport aux résultats obtenus en considérant comme correctes seulement les propositions correspondant aux équivalents trouvés dans le corpus de référence. Dans le premier cas, la précision pour les dix mots étudiés est de 71,91 % et le rappel de 68,26 %, tandis que, dans le deuxième cas, la précision est de 38,48 % et le rappel de 36,53 %. Nous remarquons aussi l'influence des limitations du corpus sur les résultats ; ceux-ci se détériorent clairement dans le cas d'équivalents qui ont une fréquence d'occurrence inférieure à dix dans le corpus d'apprentissage. Les correspondances modélisées entre le mot source et ces équivalents ne contiennent pas d'informations suffisantes sur leurs conditions d'utilisation afin qu'ils soient sélectionnés pour traduire une nouvelle occurrence du mot. Si les résultats pour ces équivalents ne sont pas pris en compte, la précision est de 80,33 % et le rappel de 76,8 % en tenant compte de la clusterisation versus 48,11 % et 46 % dans le cas inverse.

[15] Par exemple, la proposition de l'équivalent *διακίνηση* pour une occurrence de *movement* traduite dans le corpus de test par *κυκλοφορία*. Cette proposition est considérée comme correcte étant donné que les deux équivalents sont liés et qu'ils sont considérés comme adéquats et interchangeables au sein du nouveau contexte.

6. Discussion et perspectives

La méthode présentée dans cet article ne se base sur aucune ressource sémantique prédéfinie, et les sens sont mis en évidence sur la base des informations contenues dans le corpus d'apprentissage. Il serait donc souhaitable de procéder à une validation des sens proposés et des relations sémantiques établies entre les équivalents. Pour ce faire, nous avons comparé les résultats obtenus par la méthode des cooccurrences avec ceux d'une autre méthode automatique de désambiguïsation ; il s'agit de la méthode des Miroirs Sémantiques proposée par Dyvik (1998, 2003) qui se sert des informations sur les relations de traduction entre les mots d'un bitexte afin d'analyser leur sémantique et de construire des thesaurus sémantiques. En raison des contraintes d'espace, nous ne pouvons pas présenter ici en détail les résultats de l'application de cette méthode sur notre corpus. Cependant, nous pouvons signaler que les distinctions sémantiques proposées par cette méthode sont assez similaires aux distinctions proposées par la méthode contextuelle. Etant donné que les Miroirs Sémantiques utilisent des informations de nature très différente[16], la similarité entre les résultats des deux méthodes constitue une sorte de validation des résultats obtenus par la méthode contextuelle et apporte un bon degré de confiance quant aux distinctions sémantiques proposées.

L'application à notre corpus de la méthode des Miroirs Sémantiques permet aussi de remédier à une limitation de la méthode contextuelle de désambiguïsation. Celle-ci rend possible le repérage des relations de similarité sémantique entre les équivalents et leur clusterisation, mais elle ne permet pas l'analyse de la nature des relations entre les équivalents clusterisés. Les différents équivalents de traduction d'un mot peuvent entretenir des relations de (quasi)-synonymie, d'hyponymie, d'hyperonymie ou, même, de causalité. La méthode des Miroirs Sémantiques offre la possibilité d'analyse de la nature des relations sémantiques existant entre les équivalents, qui sont décrites au sein des entrées correspondantes du thesaurus.

Compte tenu, d'une part, du caractère automatique de la méthode de désambiguïsation proposée dans cet article et, d'autre part, des résultats encourageants de l'évaluation, nous estimons que cette méthode mérite d'être testée à grande échelle. La manière dont l'évaluation est menée ici (par prise en compte d'un nombre fixe et plus ou moins égal de nouvelles occurrences par candidat de traduction) ne permet pas de comparer les résultats avec une méthode « baseline », qui concernerait, par exemple, la sélection de l'équivalent (et du sens) le plus fréquent pour toutes les occurrences[17]. Par la suite, nous procéderons à une évaluation plus globale de la méthode sur l'ensemble du corpus de test, qui nous permettra aussi d'avoir des résultats significatifs à l'aide de la méthode « baseline », comparables avec les résultats de notre méthode.

Références

APIDIANAKI M. (2006) Traitement de la polysémie lexicale dans un but de traduction. Actes de *TALN 2006*, 10-13 avril, Leuven, Belgique, vol. 1, 53-62.

BROWN P. F., DELLA PIETRA S. A., DELLA PIETRA V. J., MERCER R. L. (1991) Word-sense disambiguation using statistical methods. Actes de *29th Annual Meeting of the Association for Computational Linguistics,* 264-270.

[16] Les informations contextuelles ne sont pas prises en compte.
[17] Nous estimons que la comparaison avec une méthode « baseline » de ce type serait beaucoup plus intéressante que la comparaison avec une méthode effectuant le choix aléatoire entre les alternatives.

BRUN C., JACQUEMIN B., SEGOND F. (2001) Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale, TAL, vol. 42(3), 667-690.

DAGAN I., ITAI A. (1991) Word Sense Disambiguation Using a Second Language Monolingual Corpus, Computational Linguistics, vol. 20(4), 563-596.

DUFOUR N. (1997) DEFIDIC, a lexical database for computerized translation selection, *RISHH* vol. 33, Liège, 79-111.

DYVIK H. (1998) A translational basis for semantics. *Corpora and Cross-Linguistic Research : Theory, Method and Case Studies*, Johansson S., Oksefjell S. (éds.), 51-86.

DYVIK H. (2003) Translations as a semantic knowledge source. (brouillon) URL : http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/transknow.pdf

FUCHS C. (1994) *Paraphrase et énonciation*. Paris : Editions Ophrys.

GAVRILIDOU M., LABROPOULOU P., DESIPRI E., GIOULI V., ANTONOPOULOS V., PIPERIDIS S., (2004) Building parallel corpora for eContent professionals. Actes de *MLR 2004, PostCOLING Workshop on Multilingual Linguistic Resources,* Genève, 28 août.

GREFENSTETTE G. (1994) *Explorations in Automatic Thesaurus Discovery*. Boston/Dordrecht/London : Kluwer Academic Publishers.

JOHANSSON S., OKSEFJELL S. (éds.) (1998) *Corpora and Cross-Linguistic Research : Theory, Method and Case Studies*. Amsterdam/Atlanta : Rodopi.

KAJI H. (2003) Word Sense Acquisition from Bilingual Comparable Corpora. Actes de *HLT-NAACL*, Edmonton, mai-juin, 32-39.

KOEHN P. (2003) *Europarl : a Multilingual Corpus for Evaluation of Machine Translation*. (brouillon) URL : http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl.pdf

MILLER G. A., CHARLES W.G. (1991) Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, vol. 6(1), 1-28.

PANTEL P., LIN D. (2002) Discovering Word Senses from Text. Actes de *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Edmonton, 613-619.

SCHÜTZE H. (1998) Automatic Word Sense Discrimination. *Computational Linguistics,* vol. 24(1), 97-123.

SCHMID H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. Actes de *International Conference on New Methods in Language Processing*, Manchester, 44-49.

SIMARD M., LANGLAIS P. (2003) Statistical Translation Alignment with Compositionality Constraints. Actes de *HLT-NAACL Workshop : Building and Using Parallel Texts : Data-Driven Machine Translation and Beyond,* Edmonton, Canada, May 31, 19-22.

SPECIA L., VOLPE NUNES M. DAS GRAÇAS, STEVENSON M. (2006) Translation Context Sensitive WSD, Actes de *EAMT-2006 : 11th Annual Conference of the European Association for Machine Translation*, 19-20 juin, Oslo, 227-232.

VERONIS J. (2003) Hyperlex : cartographie lexicale pour la recherche d'informations. Actes de *TALN 2003*, Batz-sur-mer, 11-14 juin, 265-274.

VICKREY D., BIEWALD L., TEYSSIER M., KOLLER D. (2005) Word-Sense Disambiguation for Machine Translation, Actes de *HLT/EMNLP*, Vancouver, BC, 771-778.

Session
Syntaxe & ressources

PrepLex : un lexique des prépositions du français pour l'analyse syntaxique

Karën FORT[1], Bruno GUILLAUME[2]
[1] projets Calligramme et TALARIS, [2] projet Calligramme
LORIA/INRIA Lorraine, UMR 7503, Nancy
{Karen.Fort,Bruno.Guillaume}@loria.fr

Résumé. PrepLex est un lexique des prépositions du français. Il contient les informations utiles à des systèmes d'analyse syntaxique. Il a été construit en comparant puis fusionnant différentes sources d'informations lexicales disponibles. Ce lexique met également en évidence les prépositions ou classes de prépositions qui apparaissent dans la définition des cadres de sous-catégorisation des ressources lexicales qui décrivent la valence des verbes.

Abstract. PrepLex is a lexicon of French prepositions which provides all the information needed for parsing. It was built by comparing and merging several authoritative lexical sources. This lexicon also shows the prepositions or classes of prepositions that appear in verbs sub-categorization frames.

Mots-clés : prépositions, lexique, analyse syntaxique.

Keywords: prepositions, lexicon, parsing.

1 Introduction

Lors de la définition de classes d'entrées lexicales en fonction de leur catégorie, il apparaît naturellement une distinction entre deux types de classes. D'une part, les classes fermées dont on peut énumérer tous les éléments de façon exhaustive ; c'est le cas par exemple des pronoms ou des déterminants. D'autre part, les classes ouvertes pour lesquelles il n'est pas possible de lister tous les éléments (ils peuvent dépendre d'un vocabulaire spécifique à un domaine par exemple) ; les quatre grandes classes ouvertes sont les noms, les verbes, les adjectifs et les adverbes. La méthodologie de construction d'un lexique doit nécessairement être adaptée en fonction de cette notion de classe.

Le statut de la classe des prépositions est plus difficile à établir. A priori, l'ensemble des prépositions peut sembler une classe fermée dont on peut énumérer les éléments ; en pratique, la comparaison de différentes sources disponibles montre qu'il n'est pas facile de déterminer de façon exhaustive la liste des prépositions. Or, celles-ci représentent plus de 14% des lemmes du français[1].

[1] voir par exemple sur un corpus journalistique : `https://www.kuleuven.be/ilt/blf/rechbaselex_kul.php\#freq` (Selva *et al.*, 2002)

Dans un lexique complet, il est important d'avoir des informations de sous-catégorisation pour les mots prédicatifs (Briscoe & Carroll, 1993; Carroll & Fang, 2004). Ces informations de sous-catégorisation font souvent référence à des prépositions dans la description des arguments. En effet, ces arguments sont souvent contraints à utiliser une préposition particulière (par exemple *compter sur*) ou un ensemble de prépositions qui ont un aspect sémantique commun (par exemple *aller LOC*, où *LOC* peut être remplacé par n'importe quelle préposition locative).

Pour une analyse syntaxique profonde, il est utile de distinguer les compléments indirects requis par le verbe des autres compléments adjoints ne figurant pas dans la valence du verbe. Les deux exemples (1a) et (1b) ont la même structure de surface et seule la sémantique permet de distinguer les deux usages différents de la préposition *sur* : elle introduit un complément oblique dans le premier cas et un complément adjoint dans le second. Ce problème peut être géré par l'utilisation d'informations sémantiques plus fines.

1a. *Jean compte sur ses amis*
1b. *Jean compte sur ses doigts*

Cette distinction amène à distinguer deux usages différents des prépositions et est donc une source d'ambiguïté lexicale. Pour limiter cette ambiguïté, il est important que le lexique repère les prépositions qui peuvent jouer ces deux rôles (ce sont les prépositions argumentales).

Notre travail est destiné à fournir un lexique utilisable par un analyseur syntaxique. Nous nous sommes restreints aux aspects purement syntaxiques et à quelques éléments sémantiques comme la définition des ensembles de prépositions ayant un aspect sémantique commun (comme *LOC*). Le lexique produit est diffusé sous licence libre et a vocation à être intégré dans des ressources plus larges, existantes ou à venir.

La section 2 décrit les sources utilisées et la méthodologie de comparaison, la section 3 décrit les résultats de cette comparaison. La section 4 décrit comment le lexique est constitué à partir des résultats précédents. Enfin, la section 5 présente un exemple d'utilisation de ce lexique en analyse syntaxique.

2 Méthodologie

L'utilisation de prépositions dans le cadre d'une analyse syntaxique nécessite une liste large, inventoriant à la fois les prépositions non argumentales et celles susceptibles d'apparaître dans les cadres de sous-catégorisation des verbes.

2.1 Utilisation de lexiques syntaxiques

Bien entendu, il existe déjà des lexiques syntaxiques, qui proposent un ensemble intéressant de prépositions. Ainsi, le Le*fff* (Sagot *et al.*, 2006) fournit une liste conséquente de prépositions, mais la partie syntaxique du lexique est encore en cours de développement, il présente donc peu de prépositions dans les cadres de sous-catégorisation des verbes. En outre, certaines prépositions du Le*fff* semblent obsolètes ou rares. Le dictionnaire français-UNL (Sérasset & Boitet, 2000) en propose également, mais sa couverture reste limitée et la qualité des entrées est encore inégale. D'autres sources proposent des prépositions dans les cadres de sous-catégorisation des verbes, mais les listes ne sont pas tout à fait cohérentes d'une source à l'autre.

Nous avons donc effectué, dans un premier temps, un travail d'inventaire des prépositions présentes dans un certain nombre de ressources, des lexiques et/ou dictionnaires d'une part, pour la liste générale, des lexiques syntaxiques d'autre part, pour la liste de prépositions argumentales. Deux ressources se classent cependant dans les deux catégories, le Le*fff* et le dictionnaire UNL :

– Le Le*fff* (Lexique des Formes Fléchies du Français (Sagot *et al.*, 2006)) est un lexique morphologique et syntaxique du français à large couverture (plus de 110 000 lemmes). Dans sa version 2.2.1, ce lexique contient 48 prépositions simples et 164 prépositions complexes. Il présente également des informations de sous-catégorisation des verbes, qui font apparaître 14 prépositions que nous qualifierons d'argumentales.
– UNL (Universal Networking Language (Sérasset & Boitet, 2000)), est un dictionnaire du français vers un anglais désambiguïsé conçu pour la traduction automatique, qui comprend des informations syntaxiques dans sa partie française. UNL n'a qu'une couverture assez faible (moins de 27 000 lemmes), mais il propose dans sa partie anglaise des informations de type sémantique que nous envisageons d'utiliser par la suite. UNL contient 48 prépositions simples dont 10 apparaissant dans les cadres de sous-catégorisation des verbes.

2.2 Utilisation de sources de référence

Nous avons ensuite complété la liste de prépositions en utilisant différentes sources construites manuellement, lexique ou dictionnaire, voire grammaire :

– Le Grevisse (Grevisse, 1997), dans sa version papier, nous a permis de vérifier certaines intuitions concernant l'obsolescence ou l'usage de certaines prépositions.
– Le TLFi (Trésor de la langue française informatisé), que nous avons consulté via l'interface du CNRTL[2], offre une liste de prépositions un peu différente des autres. Elle comporte notamment les formes *voici* et *voilà*, rarement citées dans les autres sources à notre disposition.
– Enfin, la base de prépositions PrepNet (Saint-Dizier, 2006) nous a permis de vérifier à la fois la complétude de notre liste et les informations sémantiques présentes dans certaines sources.

2.3 Utilisation de dictionnaires de valences verbales

Nous avons ensuite cherché à enrichir la liste des prépositions apparaissant dans les cadres de sous-catégorisation des verbes de Le*fff* et UNL en nous référant à deux sources traitant plus particulièrement des verbes :

– Le dictionnaire de valences des verbes du français DICOVALENCE, héritier de PROTON (van den Eynde & Mertens, 2002), dont la démarche est fondée sur l'approche pronominale. Dans sa version 1.1, ce dictionnaire donne les cadres valenciels de plus de 3700 verbes. Nous avons extrait les prépositions simples et multi-mots qu'il contient (soit plus de 40), ainsi que leurs traits sémantiques associés.
– Nous avons complété cette liste de prépositions argumentales par celle de SynLex (Gardent *et al.*, 2006), lexique syntaxique créé à partir des tables du lexique-grammaire du LADL (Gross, 1975).

[2]voir : `http://www.cnrtl.fr`

	Lexiques					Cadres de sous-catégorisation			
	Le*fff*	TLFi	Grevisse	PrepNet	UNL	Le*fff*	DV[a]	SynLex	UNL
à	X	X	X	loc		319	895	887	246
après	X	X	X	loc	X	2	12	1	
aussi					X				
avec	X	X	X	X	X	35	193	611	49
chez	X	X	X	loc	X		9		1
comme	X				X	14	11	10	3
de	X	X	X	deloc	X	310	888	1980	282
depuis	X	X	X	deloc	X		2	1	
derrière	X	X	X	loc	X		3		
devers	X	X	X						
dixit	X								
emmi		X							
entre	X	X	X	loc	X		19	4	
hormis	X	X	X	X	X				
jusque	X	X	X		X		7		
lès	X	X	X						
moyennant	X	X	X	X	X				
par	X	X	X	loc	X	3	38	73	8
parmi	X	X	X	loc	X		7	7	
passé		X			X				
selon	X	X	X	X	X		1	1	
voici		X			X				

TAB. 1 – Quelques prépositions simples dans les différentes sources

[a]DICOVALENCE

Nous avons, à partir de ces différentes sources, effectué une étude systématique de la présence de chaque préposition, de leur appartenance éventuelle à des cadres valenciels, ainsi que de certains traits sémantiques qui leur sont associés. Nous avons ensuite regroupé les prépositions qui apparaissaient à la fois en tant qu'entrée lexicale et dans les cadres de sous-catégorisation des verbes.

Il est à noter que nous avons dû restreindre notre analyse des cadres de sous-catégorisation à ceux des verbes, du fait qu'il n'existe encore à notre connaissance aucun lexique présentant une information syntaxique suffisamment riche sur les adjectifs ou les noms.

Les prépositions multi-mots présentant des caractéristiques (nombre) et des difficultés spécifiques (segmentation), nous en avons fait un inventaire séparé selon les mêmes méthodes.

	Lexiques					Cadres de sous-catégorisation			
	Le*fff*	TLFi	Grevisse	PrepNet	UNL	Le*fff*	DV[a]	SynLex	UNL
à cause de	X		X	X					
à la faveur de			X	X					
à partir de	X		X	deloc				1	
afin de	X	X	X	X					
au nord de				loc					
au vu de	X								
auprès de	X	X	X	loc			27	35	
comme de							1		
conformément à	X			X					
d'avec			X				1	6	
en faveur de	X		X	X			13		
il y a	X								
jusqu'à	X			loc	X		10		
jusqu'en	X								
jusqu'où	X								
loin de	X		X	loc					
par suite de			X						
pour comble de	X								
près de	X		X	loc					
quant à	X	X	X						
tout au long de	X			X					
vis-à-vis de	X		X	X				1	

TAB. 2 – Quelques prépositions multi-mots dans les différentes sources

[a]DICOVALENCE

3 Résultat de la comparaison des sources

3.1 Prépositions simples

Nous avons ainsi listé 85 prépositions simples, dont 24 apparaissent dans des cadres de sous-catégorisation de verbes (cf. tableau 1).

Il est à noter que les 4 sources qui décrivent des cadres de sous-catégorisation utilisent des formats très différents. Le*fff* propose une vision condensée des verbes, les cadres valenciels étant regroupés dans une seule entrée ; à l'opposé, SynLex et DICOVALENCE décrivent de nombreux cadres en distinguant par exemple systématiquement les différentes réalisations syntaxiques des arguments. Les valeurs relatives sur une ligne ne reflètent donc pas du tout la couverture lexicale des sources.

3.2 Prépositions multi-mots

Nous avons obtenu une liste de 222 prépositions multi-mots, dont 18 apparaissent dans des cadres de sous-catégorisation de verbes (cf. tableau 2). Il est intéressant de noter que seuls

DICOVALENCE et SynLex proposent des prépositions multi-mots dans leurs cadres de sous-catégorisation. Le Le*fff* fournit quant à lui une liste impressionante de prépositions multi-mots (plus de 150) qui représente une excellente base de travail.

4 Construction du lexique

Le premier critère de sélection que nous avons choisi d'appliquer pour construire notre lexique est qu'une préposition doit être listée dans au moins une source parmi celles citées. Par ailleurs, nous considérons qu'une préposition est argumentale si elle apparaît dans au moins un cadre de sous-catégorisation de verbe.

4.1 Filtrage manuel

Nous avons ensuite trié ces prépositions en fonction de critères simples et avons identifié en particulier celles qui étaient à éliminer parce que :
– visiblement erronées, c'est le cas par exemple de *aussi*, présent dans le dictionnaire UNL en tant que préposition,
– obsolètes ou d'un emploi extrêmement rare, comme *emmi* (TLFi), *devers* (Le*fff*, TLFi, Grevisse) ou encore *comme de* (DICOVALENCE).
Nous avons également effectué un tri dans les traits sémantiques et avons supprimé les entrées erronées, telles que *avec* comme locatif dans SynLex et dans DICOVALENCE.

4.2 Quelques remarques

Certaines sources font apparaître comme des prépositions des formes qui ne sont pas considérées comme telles en linguistique. C'est le cas en particulier de :
– *comme*, qui n'est pas cité dans les trois sources de référence que sont le Grevisse, le TLFi et PrepNet, il est en effet ambigu et peut également être considéré comme une conjonction,
– *il y a* ou *y compris*, qui ne sont citées que dans le Le*fff*,
– *d'avec*, qui, bien qu'il apparaisse dans le Grevisse, n'est présent que dans les cadres de sous-catégorisation de DICOVALENCE et SynLex.
Nous avons choisi de conserver ces formes dans notre lexique, pour des raisons pratiques liées à l'application visée, l'analyse syntaxique.

Par ailleurs, même si sa couverture est large, notre lexique n'est évidemment pas exhaustif. Il serait intéressant d'y ajouter certaines formes manquantes, notamment :
– des prépositions présentes dans le DAFLES (Selva *et al.*, 2002), comme par exemple la forme *au détriment de*,
– des prépositions citées dans des grammaires de référence, comme *question*, dans la Grammaire méthodique du français (Riegel *et al.*, 1997),
– les multiples prépositions locatives (et, par métonymie, temporelles) que peut préfixer la forme *jusqu'*, par exemple *jusqu'auprès de*. Cette forme élidée de *jusque* pourrait sans doute faire l'objet d'un traitement particulier en tant que *modifieur* de préposition. Il en va d'ailleurs de même de *dès*, suivi d'un temporel (ou d'un locatif, par métonymie).

Lexiques						Cadres de sous-catégorisation				
Le*fff*	TLFi	Grevisse	PrepNet	UNL	**PrepLex**	Le*fff*	DV	SynLex	UNL	**PrepLex**
44	69	55	36	46	**49**	14	24	18	11	**23**

TAB. 3 – Total des prépositions simples par source

Lexiques						Cadres de sous-catégorisation				
Le*fff*	TLFi	Grevisse	PrepNet	UNL	**PrepLex**	Le*fff*	DV	SynLex	UNL	**PrepLex**
166	11	77	89	2	**206**	0	16	4	0	**15**

TAB. 4 – Total des prépositions multi-mots par source

Ce tri a également permis de mettre en évidence certaines difficultés, en particulier les élisions dans les formes multi-mots, telle *afin de*, *afin d'*, ou les contractions telles *face à*, *face au* ou *à partir de*, *à partir du*, qui seront traitées lors de la segmentation.

D'autres, comme *lès*, qui n'est utilisé qu'en toponymie dans des formes avec tirets (comme *Bathelémont-lès-Bauzemont*), seront traitées en amont, lors de la segmentation.

4.3 Résultats

Au final, nous obtenons une liste de 49 prépositions simples, dont 23 apparaissent dans des cadres de sous-catégorisation des verbes, dans au moins une source, et sont donc considérées comme argumentales (cf. tableau 3).

Nous obtenons également une liste de plus de 200 prépositions multi-mots, dont 15 apparaissent dans des cadres de sous-catégorisation des verbes, dans au moins une source, et sont donc considérées comme argumentales (cf. tableau 4).

Nous avons pour l'instant limité les informations sémantiques utilisées dans le lexique à *loc* (locatif) et *deloc* (délocatif), mais nous avons l'intention d'étendre ces catégories à celles retenues dans DICOVALENCE (temps, quantité, manière).

En outre, nous nous sommes également référés aux sources pour renseigner les catégories des arguments introduits par les prépositions argumentales.

Il n'existe pas encore de format normalisé pour les lexiques syntaxiques, même si un effort d'homogénéisation est en cours (Projet Lexsynt). Actuellement, PrepLex est donc présenté dans un format texte qui permet à la fois l'édition manuelle et l'utilisation dans un analyseur ou dans d'autres outils de traitement de la langue. Dans ce format, les informations syntaxiques sont décrites à l'aide structures de traits récursives de profondeur 2. Le niveau externe décrit la structure en termes d'"argument" : ce niveau contient toujours un trait "head" et un trait pour chacun des "arguments" de l'entrée. Le niveau interne décrit alors plus finement chaque argument. De plus, ce format permet de définir des informations syntaxiques de façon modulaire en factorisant les parties redondantes. Dans le cas des prépositions, toutes les entrées partagent le même squelette :

```
Prep : [
head [cat=prep, prep=#, funct=#]
comp [cat=#, cpl=@]
]
```

Lorsque ce squelette est instancié pour une préposition particulière, les 3 valeurs de traits (notées "#") doivent impérativement être renseignées et la valeur de trait notée "@" est optionnelle. Ces traits sont repérés par leur nom (`prep`, `funct`) pour la tête, par une notation pointée (`comp.cat`, `comp.cpl`) pour l'argument.

```
à          Prep  [prep=a|LOC; funct=aobj|loc|adj; comp.cat=np|sinf;
                  comp.cpl=void|ceque]
après      Prep  [prep=apres|LOC; funct=obl|loc|adj; comp.cat=np]
avec       Prep  [prep=avec; funct=obl|adj; comp.cat=np]
à_travers Prep  [prep=a_travers;funct=obl|adj; comp.cat=np]
```

Techniquement, la seule difficulté est de choisir comment représenter l'appartenance à une classe sémantique de prépositions comme *loc*. Ici, nous avons choisi de définir comme valeurs atomiques possibles pour le trait "prep", l'ensemble des prépositions argumentales et l'ensemble des classes sémantiques (notées en majuscules). On utilise alors la disjonction `a|LOC` pour indiquer que la préposition *à* peut être utilisée soit comme une préposition particulière, soit comme une préposition locative.

Nous avons en outre décidé d'indiquer dans le lexique les sources dans lesquelles la préposition apparaît, afin de permettre un éventuel filtrage pour des utilisations particulières. Dans le cas des prépositions argumentales, nous avons ajouté un champ comportant la fréquence d'apparition et un exemple pris dans l'une des sources.

5 Un exemple d'utilisation dans un système TAL

Nous exposons ici de manière succinte quelques problèmes posés par les prépositions lors de l'analyse syntaxique.

5.1 Spécificités liées à la segmentation

La première difficulté pour l'intégration des prépositions dans un analyseur se situe au niveau de la segmentation en lexèmes. Il faut gérer les phénomènes d'élision : *au* doit être traité comme *à le*, le *de* finissant certaines prépositions multi-mots peut être élidé en *d'*. Cependant, ces phénomènes ne sont pas spécifiques aux prépositions, il sont traités soit dans le lexique (par exemple L*efff* distingue les deux formes *au cours de* et *au cours d'*), soit lors de la segmentation. Nous avons choisi de les traiter dans le segmenteur, afin de simplifier la maintenance du lexique.

Une difficulté plus directement liée aux prépositions multi-mots, est la possibilité d'avoir une ambiguïté de segmentation. Par exemple, dans les deux phrases (2a) et (2b) la suite de mots *au cours de* est une préposition multi-mots dans le premier cas, mais elle doit être décomposée dans le deuxième. D'autres prépositions multi-mots ne nécessitent jamais de découpage, par exemple *y compris*.

2a. *Il a beaucoup travaillé au cours de cette année*
2b. *Il a beaucoup travaillé au cours de M. Durand*

5.2 Prépositions *compléments* vs prépositions *argumentales*

Lors de l'analyse syntaxique, on doit nécessairement distinguer l'usage d'une préposition pour introduire un argument du verbe de celui d'une préposition introduisant un complément. Comme on l'a vu (exemples (1a) et (1b)), cette distinction est souvent difficile à établir et repose sur des considérations sémantiques. L'analyse syntaxique doit alors maintenir l'ambiguité de rattachement. Le fait d'avoir des informations précises sur les prépositions argumentales permet de contrôler ces ambiguïtés.

6 Conclusion

En comparant divers lexiques et dictionnaires, nous avons établi une liste de prépositions utiles pour le TAL. Nous nous sommes concentrés surtout sur les aspects syntaxiques. Un tri manuel a permis d'écarter des prépositions obsolètes ou très rares et quelques cas d'erreur. Le lexique ainsi produit contient plus de 250 prépositions dont 49 sont des prépositions simples.

Dans les lexiques syntaxiques, les cadres de sous-catégorisation décrivent les prépositions introduisant certains arguments. Des prépositions apparaissant dans les entrées verbales d'un lexique syntaxique sont appelées argumentales. Nous avons identifié 40 prépositions argumentales.

Le lexique développé est librement disponible [3]. Ce lexique va nécessairement évoluer. D'autres sources d'information auraient leur place dans ce travail, notamment les champs *constructions* des verbes du TFLi qui font référence à des prépositions qui sont donc argumentales. Il est prévu d'utiliser prochainement cette source pour faire évoluer le lexique.

La mise en place d'une base de donnée contenant ce lexique est en cours [3]. Cela devrait permettre de faciliter la maintenance du lexique, mais aussi d'enrichir les données pour chaque entrée, notamment avec des exemples d'usage ou des traits sémantiques plus variés (*loc, deloc*, mais aussi *tim, man, qty*). Nous envisageons d'y ajouter des informations de fréquence sur corpus.

Une tâche de bien plus grande ampleur serait d'enrichir ce lexique avec des informations sémantiques plus fines que la seule référence aux classes *loc, deloc, ...* Il existe de nombreux travaux de linguistique portant sur les prépositions. Cependant, la plupart s'attache à des descriptions sémantiques fines d'un petit nombre de prépositions ; une exception notable étant le travail réalisé dans PrepNet (Saint-Dizier, 2006). Il conviendrait donc de transformer ces ressources pour les rendre directement utilisables par un système de traitement automatique des langues.

Remerciements

Nous tenons à remercier chaleureusement M. Guy Perrier, Professeur à l'Université Nancy 2, pour ses conseils patients et éclairés.

[3] `http://loriatal.loria.fr/Resources.html`

Références

BRISCOE T. & CARROLL J. A. (1993). Generalised probabilistic LR parsing for unification-based grammars. *Computational Linguistics*.

CARROLL J. A. & FANG A. C. (2004). The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, p. 107–114, Sanya City, China.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2006). Extraction d'information de sous-catégorisation à partir des tables du LADL. In *Proceedings of TALN 06*, p. 139–148, Leuven.

GREVISSE M. (1997). *Le Bon Usage – Grammaire française, édition refondue par André Goosse*. Paris – Louvain-la-Neuve : DeBoeck-Duculot, 13^e edition.

GROSS M. (1975). *Méthodes en syntaxe*. Hermann.

RIEGEL M., PELLAT J.-C. & RIOUL R. (1997). *Grammaire méthodique du français*. PUF, 3^e edition.

SAGOT B., CLÉMENT L., VILLEMONTE DE LA CLERGERIE E. & BOULLIER P. (2006). The Lefff 2 syntactic lexicon for French : architecture, acquisition, use. In *Actes de LREC 06, Gênes, Italie*.

SAINT-DIZIER P. (2006). PrepNet : a Multilingual Lexical Description of Prepositions. In *LREC, Gènes, 12/05/2006-14/05/2006*, p. 877–885 : European Language Resources Association (ELRA).

SELVA T., VERLINDE S. & BINON J. (2002). Le DAFLES, un nouveau dictionnaire pour apprenants du français. In *Actes du dixième congrès EURALEX'2002 (European Association for Lexicography). Copenhague*.

SÉRASSET G. & BOITET C. (2000). On UNL as the future "html of the linguistic content" and the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter. In *Proceedings of COLING 2000, Saarebruecken, Germany*.

VAN DEN EYNDE K. & MERTENS P. (2002). *La valence : l'approche pronominale et son application au lexique verbal*, In *Journal of French Language Studies*, p. 63–104. Cambridge University Press, 13^e edition.

Comparaison du *Lexique-Grammaire* des verbes pleins et de DICOVALENCE : vers une intégration dans le L*efff*

Laurence DANLOS[1], Benoît SAGOT[2]
[1] Lattice - Université Paris 7 - Institut Universitaire de France
2 place Jussieu, case 7003, 75251 Paris Cedex 05, France
[2] Projet Signes - INRIA, Dom. Universitaire
351 cours de la Libération, 33405 Talence Cedex, France
laurence.danlos@linguist.jussieu.fr
benoit.sagot@inria.fr

Résumé. Cet article compare le *Lexique-Grammaire* des verbes pleins et DICOVALENCE, deux ressources lexicales syntaxiques pour le français développées par des linguistes depuis de nombreuses années. Nous étudions en particulier les divergences et les empiètements des modèles lexicaux sous-jacents. Puis nous présentons le L*efff*, lexique syntaxique à grande échelle pour le TAL, et son propre modèle lexical. Nous montrons que ce modèle est à même d'intégrer les informations lexicales présentes dans le *Lexique-Grammaire* et dans DICOVALENCE. Nous présentons les résultats des premiers travaux effectués en ce sens, avec pour objectif à terme la constitution d'un lexique syntaxique de référence pour le TAL.

Abstract. This paper compares the *Lexicon-Grammar* of full verbs and DICOVALENCE, two syntactic lexical resources for French developed by linguists for numerous years. We focus on differences and overlaps between both underlying lexical models. Then we introduce the L*efff*, large-coverage syntactic lexicon for NLP, and its own lexical model. We show that this model is able to integrate lexical information present in the *Lexicon-Grammar* and in DICOVALENCE. We describe the results of the first work done in this direction, the long term goal being the consitution of a high-quality syntactic lexicon for NLP.

Mots-clés : lexique syntaxique, *Lexique-Grammaire*, DICOVALENCE, L*efff*.

Keywords: syntactic lexicon, *Lexicon-Grammar*, DICOVALENCE, L*efff*.

1 Introduction

À l'heure actuelle, il existe deux grandes ressources lexicales syntaxiques pour le français développées depuis de nombreuses années dans des laboratoires de linguistique : le *Lexique-Grammaire* (Gross, 1975; Boons *et al.*, 1976a; Boons *et al.*, 1976b; Guillet & Leclère, 1992) et le dictionnaire DICOVALENCE (Van den Eynde & Mertens, 2006). L'objectif de cet article est de comparer ces deux ressources lexicales, afin d'en tirer le meilleur parti, et de l'intégrer dans le L*efff* (Lexique des Formes Fléchies du Français, (Sagot, 2006; Sagot & Danlos, 2007), lexique syntaxique destiné au TAL et en cours de développement. Nous nous limitons ici aux seules

entrées des verbes pleins, c'est-à-dire ni figés ni supports d'adjectifs ou de noms prédicatifs (le *Lexique-Grammaire* et le L*fff* comportent des entrées pour les verbes non pleins et des entrées non verbales, mais ceci n'est pas le cas de DICOVALENCE). Les objectifs des lexiques considérés, comme de tout lexique syntaxique, sont de définir, pour chaque lemme verbal donné, ses différents emplois et, pour chacun de ces emplois, son (ou ses) cadres de sous-catégorisation et les informations complémentaires qui s'y rapportent (p.ex. les informations sur le contrôle).

Nous présentons donc un travail de comparaison entre le *Lexique-Grammaire* et DICOVALENCE, ressources présentées brièvement dans les sections 2 et 3, afin de comprendre leurs points communs et leurs divergences et d'aider à leur amélioration mutuelle (Section 4). Puis nous montrons (Section 5) en quoi le modèle lexical utilisé dans le L*fff* permet de modéliser les informations présentes dans l'une ou l'autre de ces ressources, améliorant ainsi sa précision, sa couverture, et donc la qualité des outils de TAL qui l'utilisent.

2 Introduction au *Lexique-Grammaire*

Dans le *Lexique-Grammaire*, un cadre de sous-catégorisation d´un emploi de verbe plein, qui donne sa valence, est défini par deux critères de base (non hiérarchisés) :

– nombre et nature directe ou indirecte des compléments : le tableau ci-dessous résume les schémas de cadres de sous-catégorisation définis par ce critère.

$N_0\ V$	zéro complément
$N_0\ V\ N_1$ — $N_0\ V\ Prép\ N_1$	un complément
$N_0\ V\ N_1\ N_2$ — $N_0\ V\ N_1\ Prép\ N_2$	deux compléments
$N_0\ V\ N_1\ Prép\ N_2\ Prép\ N_3$ — $N_0\ V\ Prép\ N_1\ Prép\ N_2\ Prép\ N_3$	trois compléments

  Ces schémas sont ensuite affinés selon la valeur des prépositions : sont distinguées Prép = *à*, Prép = *de*, Prép = Loc, Prép = *avec* et Prép = autres.

– réalisations du sujet et des compléments ; nous employons le terme de « position » pour désigner un élément N_i, où N_0 correspond au sujet et N_i ($i > 1$) à un complément. Une position peut-être réalisée comme une complétive (notée $Que\ P$), une infinitive (notée $Vinf$) ou un groupe nominal (noté GN). Pour chaque N_i, les distributions suivantes sont celles qui sont le plus souvent retenues : $N_i = Que\ P\ /\ Vinf\ /\ GN$ ou $N_i = Vinf\ /\ GN$ ou $N_i = GN$. À ces distributions s'ajoutent le sujet impersonnel (pléonastique) réalisé par les pronoms *il* ou *ça*, soit $N_0 =$ *ilimp* ou $N_0 =$ *çaimp*.

Les différents cadres de sous-catégorisation des emplois des verbes sont structurés en *Tables*. Chaque table est définie par une *propriété définitoire*. Parmi les propriétés définitoires, on peut distinguer les propriétés de base de celles qui sont occasionnelles :

– Les propriétés définitoires basiques correspondent à l'intersection des deux critères que nous venons d'expliquer. Ainsi la propriété définitoire de la Table 9 est $N_0\ V\ (QueP)_1\ à\ N_2$, ce qui correspond au schéma $N_0\ V\ N_1\ à\ N_2$ en ce qui concerne le nombre et la nature des compléments, et au fait que la réalisation de N_1 est $N_1 = Que\ P\ /\ Vinf\ /\ GN$. Par exemple, le verbe *dire* appartient à la Table 9 parce qu'il est la tête verbale des phrases suivantes : *(Luc)*$_0$ *a dit à (Marie)*$_2$ *(qu'il faisait beau)*$_1$ */ (être malade)*$_1$ */ (une bêtise)*$_1$[1].

[1]Dans ces phrases, *à* N_2 apparaît avant N_1, en accord avec le fait que les propriétés définitoires n'imposent pas d'ordre sur les compléments.

– Les propriétés définitoires occasionnelles servent à affiner la structuration en tables obtenue par les propriétés de base. Ainsi les verbes entrant dans le schéma $N_0\ V$ avec $N_0 = GN$ sont répartis en deux tables selon le trait plus ou moins humain du sujet. On peut donc inclure des traits sémantiques sur les $N_i = GN$ dans les propriétés définitoires. On peut aussi inclure des traits morphologiques soit sur le verbe (par exemple, la table 32RA a pour propriété définitoire $N_0\ V\ N_1$ avec $N_0 = GN$, $N_1 = GN$ et V est un verbe dérivé d'un adjectif, voir *assombrir* > *sombre*), soit sur un $N_i = GN$ (par exemple, la Table 32CV a pour propriété définitoire $N_0\ V\ N_1\ en\ N_2$ avec $N_0 = GN$, $N_1 = GN$, $N_2 = GN$ avec N_2 qui est dérivé d'un verbe, voir *caraméliser* > *caramel*). Enfin, les propriétés définitoires occasionnelles peuvent inclure une relation interdite ou autorisée entre deux phrases. Par exemple, la Table 32NM a pour propriété définitoire $N_0\ V\ N_1$ avec $N_0 = GN$, $N_1 = GN$ et où la forme passive est interdite (*Cette valise pèse 10 kilos, *10 kilos sont pesés par cette valise*). La Table 34L0 a pour propriété définitoire la relation de paraphrase entre $N_0\ V\ Loc\ N_1$ et, par abus de notation, $N_1\ V\ de\ N_0$ (*Des abeilles grouillent dans le jardin, Le jardin grouille d'abeilles*).

Les propriétés définitoires basiques et occasionnelles débouchent sur 61 tables. Les critères nombre et nature des compléments et réalisation des positions n'étant pas hiérarchisés, la structuration en tables peut être présentée de deux manières : (i) la présentation classique, où la réalisation des positions est le premier critère mis en avant, qui distingue les tables à complétive ou infinitive ($\exists\ N_i$ avec $N_i \neq GN$) regroupant les Tables 1 à 18 de (Gross, 1975), des tables sans complétive ni infinitive ($\forall\ N_i,\ N_i = GN$) regroupant les Tables 30 à 40 de (Boons *et al.*, 1976a; Boons *et al.*, 1976b; Guillet & Leclère, 1992) ; (ii) la présentation de (Leclère, 1990) où le nombre et la nature des compléments est le premier critère mis en avant.

Il nous reste à présenter l'aspect matriciel des tables. Un verbe (emploi de verbe) satisfaisant la propriété définitoire d'une table est l'en-tête lexicale d'une ligne de la table. Une table comporte plusieurs colonnes qui indiquent les propriétés respectées ou non par les en-têtes lexicales de la table. Les cases de la matrice relient les propriétés aux en-têtes lexicales avec des + et des -.

3 Introduction à DICOVALENCE

DICOVALENCE est un dictionnaire de valence verbale pour le français, héritier du lexique PROTON (Van den Eynde & Mertens, 2003). Il a été développé dans le cadre méthodologique de l'Approche Pronominale — cf. par exemple (Blanche-Benveniste *et al.*, 1984). Pour identifier la valence d'un prédicat (i.e. ses dépendants et leurs caractéristiques), l'Approche Pronominale exploite la relation qui existe entre les dépendants dits *lexicalisés* (réalisés sous forme de syntagmes) et les pronoms qui couvre « en intention » ces lexicalisations possibles. Les pronoms (et les paranoms, cf. ci-dessous), contrairement aux syntagmes, aux fonctions syntaxiques ou aux rôles thématiques, ont deux avantages majeurs :

– tout en étant des éléments de référence minimale, ils sont des éléments purement linguistiques, dénués des propriétés qui rendent difficile l'interprétation de la grammaticalité d'énoncés utilisant des dépendants syntagmatiques,
– ils sont en nombre restreint : leur inventaire est fini.

La valence peut donc être obtenue sans qu'il y ait besoin d'un travail d'interprétation, à l'aide d'une vérification systématique et exhaustive des combinaisons entre les différents pronoms et le prédicat verbal. Les pronoms retenus forment un ensemble plus large que ce que l'on désigne usuellement par le terme de « pronom » : il s'agit des pronoms clitiques, des pronoms person-

nels pleins et des pronoms dits *suspensifs* (qui regroupent ce que l'on appelle habituellement pronoms interrogatifs et adverbes interrogatifs ou indéfinis, comme *à qui*, *quand*,...). Sont également pris en compte les *paranoms*, qui se distinguent des pronoms par leur modifiabilité (*rien* modifié dans *rien d'intéressant*) et l'impossibilité de reprise par un syntagme (**il ne trouve rien, les indices* vs. *il les trouve, les indices*).

Les combinaisons entre prédicats et pronoms induisent des paradigmes de portée globale. Certains correspondent à peu près aux traditionnelles fonctions syntaxiques (P0 = {*je, tu, il, elle, ..., qui, ...*} correspond à la fonction sujet (à l'exclusion du *il* impersonnel), P1 à la fonction objet direct, P2 à la fonction à-objet ou dative, etc.), d'autres permettent des distinctions plus fines que dans d'autres approches (PQ paradigme de quantité, PM paradigme de manière, etc.)[2].

DICOVALENCE proprement dit se présente comme une liste d'entrées correspondant chacune à un emploi d'un lemme verbal[3]. Sont tout d'abord donnés l'entrée et son type (prédicateur simple, verbe adjoint, verbe auxiliaire, verbe copule, verbe de dispositif, construction résultative[4]). Suivent alors les différents paradigmes qui dépendent du prédicateur (les termes de valences), avec pour chacun d'eux la liste des pronoms et paranoms qui peuvent en être la réalisation. Sont enfin indiquées certaines propriétés complémentaires, dont les passivations possibles (*passif être*, *se passif* et/ou *se faire passif*). Le tableau 1 présente l'entrée (unique) du verbe *supprimer* extraite de DICOVALENCE.

TAB. 1 – Exemple d'entrée de DICOVALENCE (repris de (Van den Eynde & Mertens, 2006)).

VAL$	supprimer : P0 P1
VERB$	SUPPRIMER/supprimer
VTYPE$	predicator simple
NUM$	80500
EX$	r : supprimer une loi / r : supprimer les obstacles
TR$	afschaffen, opheffen, intrekken, weghalen, weglaten, schrappen, doen verdwijnen
P0$	je, nous, on, qui, que, elle, il, ils, celui-ci, ceux-ci, ça
P1$	te, vous, qui, ceci, la, le, les, en Q, en, que, celui-ci, ceux-ci, ça, $se_{réfl.}$, l'un l'autre, $se_{réc.}$
RP$	passif être, se passif, se faire passif

4 Comparaison entre le *Lexique-Grammaire* et DICOVALENCE

4.1 Divergences fondamentales

Avant d'entrer dans des considérations scientifiques, signalons qu'une différence majeure entre le *Lexique-Grammaire* et DICOVALENCE réside dans le fait que DICOVALENCE est disponible librement dans son intégralité, alors que le *Lexique-Grammaire* n'est distribué que partiellement par ses dépositaires actuels[5].

[2]Un même pronom peut appartenir à plusieurs paradigmes, p.ex. *nous* appartient à P0, P1 et P2.

[3]Il y a en moyenne 2,4 entrées par lemme.

[4]On se reportera à (Van den Eynde & Mertens, 2006) pour une description précise de ces termes.

[5]Environ 60% des données sont accessibles sur le site `http://infolingu.univ-mlv.fr/`. Pour les verbes pleins, seules 41 tables (sur 61) sont distribuées. Ceci empêche d'avoir une vision globale sur les entrées d'un verbe donné, certaines entrées pouvant faire partie de tables non distribuées.

Deux stratégies différentes ont été mises en œuvre dans le développement de ces ressources. DICOVALENCE se concentre volontairement sur les verbes les plus fréquents (3738 lemmes), et, pour ces verbes, sur leurs emplois les plus fréquents (8313 entrées). *A contrario*, le *Lexique-Grammaire* s'est lancé dans une quête sans limite d'exhaustivité, à l'intérêt discutable, aboutissant à 6500 lemmes décrits par 13 375 entrées. À titre d'illustration, la Table 31H des constructions de type $N_0^{hum}\ V$ comportait 129 entrées dans l'annexe de (Boons *et al.*, 1976a) ; cette même table, qui fait partie de celles qui sont distribuées librement, comporte aujourd'hui 626 verbes, dont certains lemmes peu usités tels que *bovaryser*, *calancher*, *se curedenter*, ou *faonner*. Au niveau des entrées, la même exhaustivité est recherchée : dans la table 9 (non disponible sauf dans (Gross, 1975)), qui, rappelons-le, a pour propriété définitoire $N_0\ V\ (Que\ P)_1\ à\ N_2$, inclut des verbes tels que *bourdonner*, *bramer*, *chuinter*, *coasser*, *couiner*, *croasser*, *crépiter*, ou *gargouiller* (ces entrées sont analysées par « fusion » : *bramer* = *dire en bramant*). D'une manière plus générale, les verbes peu usités et les entrées douteuses (correspondant à des phrases traditionnellement préfixées par « ? ») sont considérés comme acceptables par le *Lexique-Grammaire*, alors que DICOVALENCE aura tendance à ne pas les conserver.

D'un point de vue méthodologique, le *Lexique-Grammaire* repose sur une structuration hiérarchique reflétée par une organisation en tables. À l'inverse, DICOVALENCE est un ensemble non structuré d'entrées. La structuration des entrées, que l'on peut formaliser par un graphe d'héritage, n'a aucune conséquence pour la réalisation d'analyseurs syntaxiques. Néanmoins, une telle structuration reflète des généralisations linguistiques pertinentes et facilite le développement et la maintenance de ressources lexicales. C'est ce qui a amené à la structuration en deux niveaux du Le*fff*, un niveau *intensionnel* structuré et un niveau *extensionnel* plat (Section 5.2)[6].

4.2 Où l'un empiète sur l'autre

Nous avons observé deux types d'empiètements entre les approches utilisées par le *Lexique-Grammaire* et par DICOVALENCE. D'une part, le *Lexique-Grammaire*, fait abondamment usage de l'approche pronominale, quoiqu'implicitement. Par exemple, le système Prép = Loc, utilisé pour les compléments prépositionnels introduits par les prépositions *à*, *de*, *dans*, *sur*, *sous*, etc., est entièrement fondé sur la pronominalisation de ces compléments par les pronoms *y*, *en*, *où*, *là*, *ici*, etc. De même, la distinction entre prépositions et complémenteurs pouvant apparaître devant les infinitives repose sur les propriétés de pronominalisation, tout comme le phénomène de chute de la préposition (et de « ce ») devant $Que\ P$. Ainsi, la table 8 regroupe des constructions du type $N_0\ V\ de\ (ce\ Que\ P)_1$ (*Luc doute de ce que Marie parte*) et du type $N_0\ V\ (Que\ P)_1$ (*Luc doute que Marie parte*), grâce à la pronominalisation en *en* et en *de cela* de la complétive, qu'elle soit ou non introduite par *de (ce)*.

D'autre part, DICOVALENCE ne se limite pas strictement à l'étude des réalisations pronominales des actants verbaux. Par exemple, le pronom *ça* peut pronominaliser une complétive à l'indicatif ou au subjonctif, mais aussi une infinitive, une concessive ou un interrogative indirecte. DICOVALENCE est donc obligé de spécifier les syntagmes dont *ça* peut être la pronominalisation. Ceci est fait par une mise entre parenthèses du type de syntagme correspondant : *ça(qpind), ça(qpsubj), ça(inf)*, etc. Le cas échéant, l'identifiant de réalisation syntagmatique est assorti d'un complémenteur : *ça(de_inf), ça(à_inf), ça(de ce qps)*, etc.

[6]Des considérations parallèles au niveau des grammaires ont induit le développement de la notion de *méta-grammaire*, description grammaticale structurée et factorisée, permettant la génération de grammaires classiques.

4.3 Où l'un apporte à l'autre

Le *Lexique-Grammaire* et DICOVALENCE sont deux ressources très riches mais incomplètes. Toutefois, elles peuvent mutuellement s'enrichir. Ainsi, DICOVALENCE comporte des informations précises sur les pronoms suspensifs (cf. plus haut), tandis que le *Lexique-Grammaire* ne prend pas en compte les interrogatives indirectes[7] et regroupent sous l'identifiant ADV les compléments pronominalisables en *ainsi, autant, (Prép) quand*, etc. A l'inverse, le *Lexique-Grammaire* comporte des informations plus précises que DICOVALENCE, en particulier :

– le système des noms parties du corps (N^{pc}), qui permet de rendre compte d'alternances telles que *Luc caresse les cheveux de Marie / Luc lui caresse les cheveux*,
– les compléments sous-catégorisés mais non pronominalisables (cf. Table 38PL : *couper le gâteau en quatre parts égales*, où le complément introduit par *en* n'est pas pronominalisable),
– certaines « restructurations », telles que *Luc copie les habitudes de Léa / Luc copie Léa dans ses habitudes*,
– certaines relations de dérivation morphologique : rappelons, par exemple, que la table 32RA regroupe les verbes morphologiquement dérivés d'adjectifs (*assombrir* $>$ *sombre*).

Certaines propriétés ou constructions complexes, comme les différentes constructions pronominales, font l'objet de codages différents dans chacune des ressources. Dans un futur proche, nous chercherons à comprendre dans quelle mesure elles s'harmonisent et/ou se complètent.

En conclusion, ces deux ressources doivent impérativement être harmonisées et mutuellement enrichies pour obtenir une ressource lexicale complète du système verbal français. Nous nous sommes attaqués à cette tâche, avec l'objectif d'enrichir une ressource destinée au TAL. C'est ce que nous allons décrire dans la section suivante, consacrée au lexique Le*fff*.

5 Le Le*fff* : présentation et enrichissement

5.1 Historique

Le développement du Le*fff* a commencé en 2003, à partir du constat suivant : à cette époque, il n'existait pas de lexique syntaxique pour le français qui soit librement utilisable et dont la couverture soit importante. La construction d'un tel lexique a donc été initiée au sein du projet Atoll de l'INRIA par Lionel Clément, avec le double objectif qu'il soit adapté au TAL tout en restant linguistiquement pertinent.

Dans un premier temps, le Le*fff* s'est limité à un lexique morphologique des verbes du français, acquis automatiquement et validé manuellement selon une technique originale (Clément *et al.*, 2004; Sagot, 2005). C'est le Le*fff* 1, distribué depuis 2004. Dans un second temps, le Le*fff* a été étendu à l'ensemble des catégories, tout en devenant un lexique morphologique *et* syntaxique. L'extension à toutes les catégories a été faite manuellement pour les catégories fermées, et à l'aide du lexique morphologique français de Multext (Veronis, 1998) pour les noms, adjectifs et adverbes — lexique dont la libre exploitation nous a été autorisée explicitement par son principal auteur. Les informations syntaxiques ont été tout d'abord renseignées intégralement à la main, en profitant au mieux de l'architecture à deux niveaux du Le*fff* (cf. ci-dessous). Depuis, diverses techniques ont été utilisées pour étendre et corriger le Le*fff* : acquisition automatique (avec validation manuelle) d'entrées morphologiques et d'informations syntaxiques atomiques

[7] Voir cependant (Nakamura, 2006), qui a codé les interrogatives indirectes pour la Table 6 ($N_0\ V\ (Que\ P)_1$).

(Sagot, 2006; Sagot *et al.*, 2006), corrections et ajouts manuels ou guidés par des techniques automatiques telles la fouille d'erreurs dans les sorties d'analyseurs syntaxiques (Sagot & Villemonte de La Clergerie, 2006), et recherche de mots inconnus dans de grands corpus.

Le L*efff* est donc aujourd'hui un lexique syntaxique à large couverture pour le français. Actuellement en version 2.5, il est entièrement téléchargeable sous sa forme extensionnelle, et sera prochainement téléchargeable également sous sa forme intensionnelle, sous une licence libre (LGPL-LR), sur le site internet `www.lefff.net`.

5.2 Modélisation des informations syntaxiques

Le L*efff* repose sur une architecture à deux niveaux : (i) un *lexique intensionnel*, où l'information est factorisée au maximum, qui associe à chaque lemme une classe morphologique et une classe syntaxique ; c'est à ce niveau qu'est fait le travail de développement — (ii) un *lexique extensionnel*, obtenu à partir du lexique intensionnel par compilation, qui associe à chaque forme une structure représentant explicitement toutes les informations linguistiques associées ; c'est ce lexique qui est utilisé par les analyseurs. Ci-dessous une entrée intensionnelle pour le lemme *manger* et une entrée extensionnelle pour la forme fléchie *mange* :

manger v-er:std @verbe_transitif_direct,

mange v [pred='manger$_1$<Suj:snlcln,Obj:(snlcla)>', cat=v, @pers, @PS13s].

Au niveau intensionnel, les informations syntaxiques sont donc décrites à l'aide de classes syntaxiques, définies par héritage de propriétés syntaxiques atomiques, propriétés elles-mêmes définies de façon indépendante de la définition des classes. Au niveau extensionnel, le cadre de sous-catégorisation d'une forme donnée[8] est constitué d'une liste de *fonctions syntaxiques*, chacune indiquant les *réalisations* possibles de cette fonction ainsi que le caractère obligatoire ou non de sa réalisation (indiqué par des parenthèses). La structure syntaxique complète, outre le cadre, comporte le cas échéant des *macros* (introduites par « @ ») qui représentent de façon implicite des informations syntaxiques complémentaires (contrôle, attribution, (im)personnel,...).

Les fonctions syntaxiques ne sont utilisées ni dans le *Lexique-Grammaire* ni dans DICOVALENCE, mais les notions respectives de position et de paradigme s'en rapprochent. Elles sont définies dans le L*efff* par des critères proches de ceux de DICOVALENCE, développés au cours de travaux de comparaison et fusion avec le *Lexique-Grammaire* et DICOVALENCE (en particulier sur les constructions impersonnelles, cf. (Sagot & Danlos, 2007)), mais également au cours de travaux non publiés en collaboration avec Claire Gardent. Ces critères reposent sur la substituabilité (en prenant en compte pronoms *et* syntagmes), sur le principe de réalisation unique d'une fonction syntaxique pour un prédicat donné, et sur l'identification de la fonction par un paradigme de pronoms (à l'exception des cas à partage d'arguments, c'est-à-dire les attributs)[9].

[8] Le passage de la forme intensionnelle à la forme extensionnelle est également un passage d'un lexique de lemmes à un lexique de formes. Ceci permet à certaines formes d'indiquer des modifications dans leur structure syntaxique par rapport à la structure par défaut correspondant à la classe syntaxique. Ainsi, bien que le conjugueur ne construise qu'une forme pour le participe passé d'un verbe passivable, le lexique extensionnel comportera deux entrées pour cette forme, l'une, active, avec le cadre par défaut, et l'autre, passive, dont les arguments syntaxiques seront caractéristiques de la construction passive.

[9] Actuellement, la liste de fonctions utilisées est la suivante : Suj (fonction sujet), Obj (fonction objet), Objà (fonction à-objet), Objde (fonction de-objet), Loc (fonction locative), Dloc (fonction délocative), Att (fonction attributive), Obl et Obl2 (fonctions obliques). Une terminologie assez traditionnelle a été préférée, pour des questions de lisibilité, à une terminologie plus algébrique comme utilisée dans DICOVALENCE. Ce qui ne signifie évidemment pas, par exemple, que toute réalisation d'un Objde comporte la préposition *de*, ni, à l'inverse, que tout complément

Les réalisations possibles, quant à elles, sont de trois types : *pronoms clitiques* (clitique nominatif (cln), accusatif (cla), datif (cld), génitif (en), locatif (y)[10]), *syntagme direct* (syntagme nominal (sn), adjectival (sa), infinitif (sinf), phrastique fini (scompl), interrogative indirecte (qcompl)) et *syntagme prépositionnel* (syntagme direct précédé d'une préposition, comme de-sn, à-sinf ou pour-sa ; à-scompl et de-scompl représentent les réalisations en *à/de ce que P*).

Le L*efff* extensionnel est illustré dans le Tableau 2. On notera que les listes de fonctions syntaxiques dans les entrées active et passive de *mangé* sont présentées en ordre inverse. Ceci pour simuler d'une façon (trop ?) économique les rôles thématiques. Nous envisageons d'indiquer plus explicitement les rôles thématiques.

TAB. 2 – Quelques entrées du L*efff* extensionnel

apprend	v	[pred='apprendre$_2$<Suj :sn|cln,Obj :(sn|cla|à-sinf|scompl|qcompl)>', cat=v, @pers, @P13s] # *Pierre apprend à conduire / la conduite*
imagine	v	[pred='imaginer$_1$<Suj :sn|cln,Obj :(sn|cla), Att :(sn|sa|sinf|comme-sn|comme-sa)>', cat=v, @pers, @PS13s] # *Pierre imagine Marie nue / se dévêtir*
mangé	v	[pred='manger$_1$<Suj :sn|cln,Obj :(sn|cla)>', cat=v, @active, @avoir, @Kms]
mangé	v	[pred='manger$_1$<Obl :(par-sn),Suj :sn|cln>', cat=v, @passive, @Kms]

5.3 Enrichir du L*efff* à partir du *Lexique-Grammaire* et de DICOVALENCE

Le L*efff* repose ainsi sur une architecture efficace et un format directement utilisable en TAL. De plus, ce format est en partie le résultat d'un consensus issu de travaux réalisés dans le cadre du projet ILF LexSynt. Nous avons donc commencé à enrichir le L*efff* à partir du *Lexique-Grammaire* et de DICOVALENCE, ce qui demande une connaissance approfondie de ces deux ressources, étant donné les divergences entre les modèles lexicaux sous-jacents.

Il n'est pas possible de convertir directement au format L*efff* les informations lexicales présentes dans le *Lexique-Grammaire*. Les travaux de (Gardent *et al.*, 2005) effectuent une telle conversion pour les tables distribuées (débouchant sur le lexique Synlex-L*efff*), mais elle est indirecte, imparfaite, et nécessite une explicitation formalisée de données linguistiques sous-entendues dans le *Lexique-Grammaire*, apport qui n'est ni simple ni aisé. Nous avons donc préféré pour le moment nous focaliser sur certaines constructions, mal traitées dans le L*efff*, et extraire du *Lexique-Grammaire* les informations lexicales pertinentes pour les y intégrer. Dans sa version actuelle (2.5), le L*efff* a été amélioré de cette façon :

– pour les constructions impersonnelles verbales et adjectivales (Sagot & Danlos, 2007), extraites de l'outil ILIMP (Danlos, 2005) développé en partie à partir du *Lexique-Grammaire*,
– pour un certain nombre d'expressions verbales figées (Danlos *et al.*, 2006).

En revanche, la conversion de DICOVALENCE au format L*efff* est plus directe. La convergence sur un nombre important de points entre les paradigmes de DICOVALENCE et les fonctions syntaxiques du L*efff* rendent la correspondance relativement simple à implémenter[11]. Les pa-

introduit par *de* est la réalisation d'une fonction Objde. Pour une description plus précise, voir (Sagot & Danlos, 2007).

[10]On notera que le *se* réfléchi ou réciproque est considéré comme une réalisation de type cla ou cld selon les cas (*Les époux se disputent / Pierre se laisse cette possibilité*).

[11]Les correspondances, à quelques exceptions près (sujet des impersonnelles, par exemple), sont les suivantes : P0 → Suj, P1 → Obj, P2 → Objà, P3 → Objde, PL → Loc, PDL → Dloc, PMi → Att, PQ → Att (discutable), PP → Obl ou Obl2, PM ignoré (pour le moment).

radigmes de pronoms de DICOVALENCE peuvent également se convertir directement en listes de réalisations au sens du Le*fff*. Toutefois, le Le*fff* ne retranscrit pas ces paradigmes dans toute leur richesse, et de l'information est donc perdue. Elle pourrait ne pas l'être si l'on prenait également en compte les traits sémantiques que l'on peut extraire des paradigmes de pronoms (Mertens, comm. pers.). Mais pour l'instant, DICOVALENCE n'a été utilisé que pour procéder à une évaluation du Le*fff*, pas pour compléter et corriger ce dernier.

5.4 Exemple d'évaluation : les constructions verbales impersonnelles

L'évaluation d'un lexique comme le Le*fff* peut se faire naturellement via l'évaluation d'analyseurs qui reposent sur lui, travail que nous effectuerons dans le futur. Mais une évaluation directe par comparaison avec d'autres ressources est également riche d'enseignements. Pour illustrer le dialogue que nous avons instauré entre le *Lexique-Grammaire*, DICOVALENCE et le Le*fff*, nous avons procédé à une comparaison entre le Le*fff* et DICOVALENCE pour les entrées verbales impersonnelles, renseignées dans le Le*fff* à partir d'ILIMP, et donc indirectement à partir du *Lexique-Grammaire*. L'évaluation s'est faite sur les entrées défactorisées (une entrée comportant des disjonctions - sur les réalisations ou à cause de fonctions syntaxiques facultatives - est remplacée par un ensemble d'entrées). On peut alors comparer les cadres complets, ou se restreindre aux cadres fonctionnels (liste des fonctions syntaxiques réalisées).

Au niveau des cadres fonctionnels, les résultats sur les constructions verbales impersonnelles sont les suivants : 60 cadres présents à la fois dans DICOVALENCE et dans le Le*fff*, 160 cadres (tous corrects) présents seulement dans le Le*fff*, et 19 cadres présents uniquement dans DICOVALENCE (certains d'entre eux nous ont semblé inacceptables, certains autres sont la conséquence de difficultés de conversion entre DICOVALENCE et le modèle lexical du Le*fff*). Le Le*fff* est donc désormais couvrant et précis sur les impersonnelles.

6 Conclusion

L'amélioration du Le*fff* à partir du *Lexique-Grammaire* et de DICOVALENCE est donc en cours dans le but d'obtenir une ressource de référence pour le TAL[12]. En ce qui concerne le *Lexique-Grammaire*, la pertinence et l'utilité de ces travaux reste limitée, compte tenu de la disponibilité restreinte des tables. Ceci n'est pas le cas de DICOVALENCE et nous comptons effectuer à partir de cette ressource, en collaboration avec Piet Mertens, un travail sur les constructions verbales pronominales pour augmenter la qualité du Le*fff* sur ce point.

En amont des données lexicales proprement dites, le modèle lexical du Le*fff* est à améliorer : certains cas verbaux non triviaux sont à revoir (en particulier, les modaux et *verbes adjoints* (selon la terminologie de DICOVALENCE), qui correspondent respectivement aux verbes de la Table 1 et aux verbes de perception de la Table 6 du *Lexique-Grammaire*), le modèle utilisé pour les constructions figées est trop simpliste, certaines fonctions syntaxiques sont probablement à découpler (distinguer Att d'un pseudo-objet de type PQ) ou à ajouter (cf. PM), et certaines réalisations, comme les concessives, sont à mieux prendre en compte.

Enfin, nous allons mettre l'accent sur différents types d'interface de visualisation et d'édition du Le*fff*, voire de comparaison avec d'autres ressources converties au format Le*fff*.

[12]Des travaux de comparaison avec Synlex-Le*fff* sont aussi en cours, en collaboration avec Claire Gardent.

Références

BLANCHE-BENVENISTE C., DELOFEU J., STEFANINI J. & EYNDE K. V. D. (1984). *Pronom et syntaxe. L'approche pronominale et son application au français*. Paris : SELAF.

BOONS J.-P., GUILLET A. & LECLÈRE C. (1976a). *La structure des phrases simples en français, Classes de constructions transitives*. Rapport interne, LADL, CNRS, Paris 7.

BOONS J.-P., GUILLET A. & LECLÈRE C. (1976b). *La structure des phrases simples en français, Constructions intransitives*. Genève : Droz.

CLÉMENT L., SAGOT B. & LANG B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proc. of LREC'04*, p. 1841–1844, Lisbon, Portugal.

DANLOS L. (2005). ILIMP : Outil pour repérer les occurences du pronom impersonnel *il*. In *Actes de TALN 2005*, Dourdan, France.

DANLOS L., SAGOT B. & SALMON-ALT S. (2006). French frozen verbal expressions : from lexicon-grammar to NLP applications. In *Actes du colloque sur le lexique et la grammaire 2006*, Palerme, Italie.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2005). Maurice Gross' grammar lexicon and natural language processing. In *Proc. of the 2nd LTC*, Poznań, Poland.

GROSS M. (1975). *Méthodes en syntaxe*. Paris : Hermann.

GUILLET A. & LECLÈRE C. (1992). *La structure des phrases simples en français : Les constructions transitives locatives*. Genève : Droz.

LECLÈRE C. (1990). Organisation du lexique-grammaire des verbes français. *Langue Française*, **87**.

NAKAMURA T. (2006). *Lexique et grammaire des interrogatives partielles en français : étude des verbes à une complétive directe*. PhD thesis, Université de Marne-la-Vallée.

SAGOT B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Proc. of TSD 2005 (LNAI 3658, ©Springer-Verlag)*, Karlovy Vary, Czech Rep.

SAGOT B. (2006). *Analyse automatique du français : lexiques, formalismes, analyseurs*. PhD thesis, Université Paris 7.

SAGOT B., CLÉMENT L., VILLEMONTE DE LA CLERGERIE E. & BOULLIER P. (2006). The Lefff 2 syntactic lexicon for French : architecture, acquisition, use. In *Proc. of LREC'06*.

SAGOT B. & DANLOS L. (2007). Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – Constructions impersonnelles et expressions verbales figées. *Cahiers du Cental*. to appear.

SAGOT B. & VILLEMONTE DE LA CLERGERIE E. (2006). Error mining in parsing results. In *Proc. of ACL 2006*, p. 329–336, Sydney, Australia.

VAN DEN EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, **13**, 63–104.

VAN DEN EYNDE K. & MERTENS P. (2006). Le dictionnaire de valence DICOVALENCE : manuel d'utilisation. http ://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf.

VERONIS J. (1998). *Multext-Lexicons, A set of Electronic Lexicons for European Languages*. Rapport interne.

Dictionnaires électroniques et étiquetage syntactico-sémantique

Pierre-André BUVET, Emmanuel CARTIER, Fabrice ISSAC, Salah MEJRI
LDI UMR 7187– Université Paris 13
{prénom.nom}@lli.univ-paris13.fr

Résumé. Nous présentons dans cet article le prototype d'un système d'étiquetage syntactico-sémantique des mots qui utilise comme principales ressources linguistiques différents dictionnaires du laboratoire *Lexiques, Dictionnaires, Informatique* (LDI). Dans un premier temps, nous mentionnons des travaux sur le même sujet. Dans un deuxième temps, nous faisons la présentation générale du système. Dans un troisième temps, nous exposons les principales caractéristiques des dictionnaires syntactico-sémantiques utilisés. Dans un quatrième temps, nous détaillons un exemple de traitement.

Abstract. We present in this paper a syntactico-semantics tagger prototype which uses as first linguistic resources various dictionaries elaborated at LDI. First, we mention several related works. Second, we present the overall sketch of the system. Third, we expose the main characteristics of the syntactico-semantic dictionaries implied in the processes. Last, using an example, we explicit the main stages of the analysis.

Mots-clés : étiqueteur sémantique, dictionnaire, LMF, XML, XPATH;

Keywords : word sense disambiguation (WSD), dictionnary, LMF, XML, XPATH;

1 Introduction

L'une des activités majeures du Laboratoire *Lexiques, Dictionnaires, Informatique* (LDI) est d'élaborer des dictionnaires électroniques à large couverture qui sont dédiés à des systèmes opérant sur des textes numérisés. Les descriptions contenues dans les dictionnaires sont de nature morphologique, d'une part, syntactico-sémantique, d'autre part. Les descriptions du second type sont effectuées dans le cadre théorique du modèle des classes d'objets (Gross, 1995, Le Pesant et Mathieu-Colas, 1998). Nous présentons dans cet article le prototype d'un système d'étiquetage syntactico-sémantique des mots qui utilise comme principales ressources linguistiques différents dictionnaires du LDI. Dans un premier temps, nous mentionnons des travaux sur le même sujet. Dans un deuxième temps, nous faisons la présentation générale du système. Dans un troisième temps, nous exposons les principales caractéristiques des dictionnaires syntactico-sémantiques utilisés. Dans un quatrième temps, nous détaillons un exemple de traitement.

2 État de l'art[1]

Nous rappelons comment l'étiquetage sémantique est généralement défini, puis nous indiquons dans quel cadre ce type d'étiquetage peut être utilisé et nous précisons finalement ses principales caractéristiques.

L'étiquetage sémantique est la tâche qui consiste à attribuer une valeur sémantique à un mot lexical[2]. Il s'agit d'une tâche intermédiaire dans un processus de traitement automatique des langues, puisqu'elle sert de point de départ à d'autres tâches plus directement en rapport avec la finalité du processus. Les étiquetages morphosyntaxiques et syntaxiques sont également des tâches intermédiaires que l'on considère disjointes de l'étiquetage sémantique.

La notion de valeur sémantique est beaucoup moins précise que celles de valeur morphosyntaxique et de valeur syntaxique car la structure et l'étendue des informations de nature sémantique sont beaucoup plus complexes. A l'instar de l'étiquetage morphosyntaxique, l'étiquetage sémantique implique que le choix d'une étiquette ne dépend pas seulement du mot mais aussi de son contexte. Ainsi dans les phrases *La pièce est dans le porte-monnaie* et *Le porte-monnaie est dans la pièce*[3], le calcul du sens du mot *pièce* nécessite des connaissances sur sa combinatoire compte tenu de sa polysémie.

Selon les types d'applications, les étiquettes sémantiques ont plus ou moins d'importance. Elles sont cruciales en traduction automatique ou en traduction assistée par ordinateur du fait que la transposition d'un texte d'une langue cible vers une langue source nécessite, entre autres tâches, d'attribuer la valeur exacte d'une forme donnée dans un contexte donné. Ainsi, pour traduire *abattre* par *cut down* dans *abattre un arbre* et *kill* par *tuer* dans *abattre un criminel*, il faut étiqueter la forme verbale comme un synonyme de *couper* dans un cas, de *tuer* dans l'autre. De même, en recherche d'information, il faut faire appel non seulement à l'étiquetage morphosyntaxique mais aussi à l'étiquetage sémantique pour construire des index et analyser les requêtes. Ainsi, une requête comme *Quelle est la vitesse du jaguar ?* ne sera pas associée à des résultats pertinents si la nature du mot *jaguar* (félin, avion ou voiture) n'est pas précisée.

L'étiquetage sémantique est conçu comme un processus en deux étapes : (i) utiliser une ressource linguistique du type dictionnaire ou du type ontologie pour attribuer un ensemble de sens à tous les mots pleins d'un texte ; (ii) utiliser des techniques symboliques ou numériques faisant appel à des ressources linguistiques pour éliminer les sens incorrects. Ce type de représentation permet de distinguer les ressources linguistiques à utiliser et les traitements à appliquer. La première étape ne présente aucune difficulté particulière pour peu que l'on dispose d'une ressource suffisamment complète ; elle consiste à attribuer des étiquettes sémantiques à des formes. La deuxième étape fait appel à des ressources linguistiques beaucoup plus riches[4] ou bien à un échantillon pré-étiqueté.

Le système utilise dans les deux cas des ressources linguistiques du type dictionnaire pour étiqueter sémantiquement les mots lexicaux. Nous montrons maintenant l'intérêt de faire appel à des dictionnaires électroniques paramétrés de façon syntactico-sémantique.

3 Présentation générale du système

L'élaboration du système est un projet en cours qui s'inscrit dans le programme TAL du LDI visant à construire une plateforme d'analyse syntactico-sémantique des textes dédiée à la mise

[1] On trouvera une description plus complète du domaine dans (Ide et Véronis, 1998) et (Véronis, 2004) ainsi qu'un aperçu des applications les plus récentes dans (Mihalcea et Edmonds, 2004).

[2] En anglais, Word Sense Disambiguation (WSD).

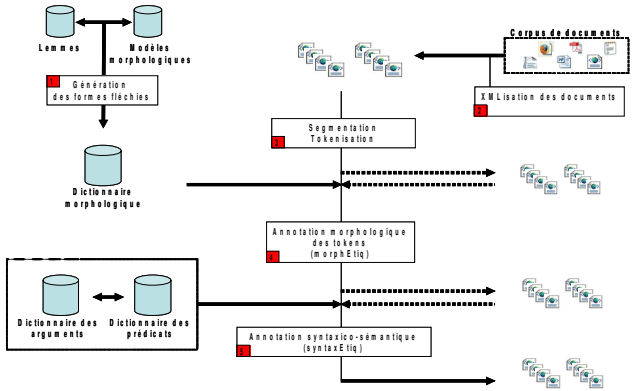
[3] Adaptation au français des phrases « the box is in the pen » et « the pen is in the box » de Yehoshua Bar-Hillel (Sabah, 1996).

[4] Par exemple, WordNet / EuroWordNet (Preiss, Stevenson, 2004) ou EST (Piao et al., 2005)

en place d'applications (veille, traduction, etc.) et à la gestion de corpus de documents par les linguistes du laboratoire .

Nous indiquons quelle est l'architecture générale du système puis nous précisons pour chaque module quelles sont les ressources utilisées. Ensuite, nous présentons le corpus de travail et nous détaillons la chaîne de traitements

3.1 Architecture générale



3.2 Ressources du système

Le système utilise trois **ressources** lexicographiques et un corpus :

- **dictionnaire morphologique** : Il comprend pour l'instant l'ensemble des formes simples du français (actuellement plus de 750 000 formes), qui sont générées à partir d'une table des lemmes et d'une table de modèles morphologiques. Par la suite, il intègrera l'ensemble des mots composés du français[5].
- **dictionnaire des arguments** : *cf. Infra.*
- **dictionnaire des prédicats** : *cf. Infra.*
- **Corpus** : il est composé pour l'essentiel d'articles du journal *Le Monde* au format texte ainsi que d'autres types de documents dans des formats variés (HTML, Open Office, PDF, Word et RTF)

Un pré-traitement permet de normaliser les documents en les convertissant de leur format d'origine vers un format XML commun. Il consiste à récupérer des informations sur la structure des documents. Nous avons opté, en l'état actuel, pour une DTD « minimale », permettant d'annoter les titres, sous-titres, sections et sous-sections ainsi que les paragraphes du texte. Dans une phase ultérieure, des annotations supplémentaires permettront de rendre compte de structures textuelles plus fines, comme les listes et les tableaux. La normalisation XML des documents permet de les décorer de balises `<text><title level='1..N'><section level='1..N'><p>`. Le pré-traitement permet également de convertir les documents de leur encodage d'origine vers l'encodage UTF-8[6].

[5] Ce dictionnaire est issu des travaux de Michel Mathieu-Colas au LLI

[6] Ce travail a été effectué avec la promotion 2006-2007 des étudiants du Master PRO TILDE de l'Université Paris 13.

Le système comprend deux étapes de pré-traitement (génération du dictionnaire morphologique, étape 1, et normalisation XML des documents, étape 2) et trois étapes de traitement. A chaque étape, le système prévoit une sortie afin d'effectuer une évaluation des résultats.

3.3 Étapes du traitement

Etape 3 : segmentation en phrases et en mots des documents XML : la segmentation en phrases et en mots permet de décorer les documents de deux nouvelles balises : `<s>` et `<w>` (respectivement « sentence » et « word »).
La phase de segmentation en phrases (définies du point de vue graphique) résulte de l'analyse et du classement des différents sortes de 'point' dans le corpus « Le Monde ». Des listes d'abréviations et d'acronymes courants ainsi qu'une série d'expressions régulières permettent de repérer les 'points' qui ne sont pas en fin de phrase. Un sous-corpus a été manuellement annoté pour constituer un corpus de référence de 150 000 mots. Les résultats de la segmentation automatique sont un taux de rappel de 98,7% et un taux de précision de 97,6%.
La phase de segmentation en mots a consisté tout d'abord à reconnaître un certain nombre d'unités textuelles, qu'il s'agisse d'entités numériques (*5,7%*, *13,2 millions*), d'entités temporelles à base numérique (*le 12/12/2005*) ou encore d'entités spécifiques (url, mail, etc.). Elle a également consisté à normaliser les signes de ponctuation (réduction de plusieurs espaces en un seul, séparation des signes de ponctuation, normalisation des signes d'élision...), puis à segmenter les phrases en mots sur la base de l'espace.[7]
Etape 4 : annotation morpho-syntaxique des tokens : les chaînes de caractères reconnues comme des mots font l'objet d'une annotation morphosyntaxique. Par projection, les informations contenues dans le dictionnaire morphologique sont associées aux tokens : lemme, catégorie morphosyntaxique, informations de nombre, genre, mode, temps, personne[8].
A ce stade, les mots peuvent être ambigus. Par exemple, *porte* comprend l'ensemble des informations liées aux formes verbales (mode : indicatif, subjonctif ; temps : présent, personne$1 : 1, 3) et à la forme nominale (nom, féminin, singulier). Dans le fichier XML adéquat, cela se traduit par l'annotation suivante :

```
<w                                                              case='lc'>
 <morph  lemma='porter'  cat='v'  mood=''subj'  tense='pres'  pers='1'/>
 <morph  lemma='porter'   cat='v'  mood=''subj'  tense='pres'  pers='3'/>
 <morph  lemma='porter'   cat='v'  mood=''ind'  tense='pres'  pers='1'/>
 <morph  lemma='porter'   cat='v'  mood=''ind'  tense='pres'  pers='3'/>
 <morph      lemma='porte'       cat='n'      nb='plu'      gender='fem'/>
 porte
</w>
```

Etape 5 : annotation syntaxico-sémantique : l'annotation syntaxico-sémantique est l'étape la plus complexe du traitement. Elle fait appel au dictionnaire des arguments et à celui des prédicats. Elle se décompose en plusieurs sous étapes :

- 5a. désambiguisation morpho-syntaxique : le système élimine un certain nombre de solutions morphologiques issues de la projection des dictionnaires sur les mots « hors contexte » en faisant appel à des règles correspondant à des grammaires locales.

[7] Dès cette phase, un certain nombre d'informations typographiques sont incluses sous forme d'attributs : le type de l'élément (valeur de l'élément : w pour word, num pour infos numériques, punct pour signe de ponctuation, symb pour symbole) ainsi que des attributs de casse pour w (case='lowercase,uppercase,titlecase') et un attribut type pour num (afin de reconnaître, entre autres, les dates).

[8] *Cf.* Mathieu-Colas, 2007, à propos des informations morphologiques utilisées.

- 5b. projection du dictionnaire des arguments : le dictionnaire des arguments est projeté sur le fichier XML afin d'ajouter des informations de type sémantique au niveau de certains signes de type mot « w ».
- 5c. Annotation des constructions syntaxiques : le système identifie diverses constructions syntaxiques : groupes nominaux, groupes prépositionnels, etc.. Pour chaque groupe, un attribut « head » correspondant à la « tête sémantique» des différents groupes est également récupéré. Sont également reconnues les séquences relatives à des entités nommées (personnes, organisations, lieux, artefacts, événements).
- 5d. Analyse prédicative : il s'agit enfin d'identifier les structures prédicat-arguments à l'aide du dictionnaire des prédicats. Le système génère automatiquement une expression XPATH à partir des descriptions lexicographiques. Par exemple, pour l'entrée « subir », dans le sens de subir(EVENEMENT,ENTITE CONCRETE), nous aurons l'expression XPATH :
  `group[1][@cat='gn' and @sem='ENTITE'] and group[2][@cat='gv' and @head='subir'] and group[3][@cat='gn' and @sem='EVE'].`

Le XML résultant de l'analyse prédicative se présente comme suit :

```
<s>
       <gn head='12/12/2003'>
              <w case='tc' ><morph lemma='le' cat='det'>Le</morph></w>
              <num type='date'>12/12/2003</num>
       </gn>
              <punct>,</punct>
<pred sem='subir' X0='population' X1='hausse'>
       <gn head='population'>
          ...
      </gn>
      <gv head='subir'>
       ...
      </gv>
      <gn head='hausse'>
       ...
     </gn>
</pred>
              <w typo='punct'>,</w>

    <gadj head='atteindre'>
          ...
      </gadj>
              <punct>.</punct>
</s>
```

4 Les dictionnaires syntactico-sémantiques

Le modèle des classes d'objets subdivise les unités linguistiques à partir du postulat suivant : toute phrase élémentaire est constituée d'un prédicat du premier ordre et de ses éventuels arguments, les autres constituants phrastiques ressortissant à l'actualisation. Il s'ensuit trois sous-catégories majeures : celle des prédicats, celle des arguments élémentaires et celles des actualisateurs. La description des items de la dernière sous-catégorie est subordonnée à celles des items des autres sous-catégories (Buvet, à paraitre). Les descriptions des arguments élémentaires et des prédicats sont formalisées dans des dictionnaires électroniques. Nous discutons de ces dictionnaires du point de vue linguistique puis du point de vue informatique.

4.1 Structuration linguistique

Deux sortes de dictionnaires syntactico-sémantiques sont élaborés au LLI : ARGU-DIC et PREDI-DIC. Le premier dictionnaire décrit les arguments élémentaires, le second les prédicats.

ARGU-DIC : les arguments élémentaires sont des substantifs qui ne peuvent jamais occuper une position prédicative dans une construction à support (Gross et Vives, 1986). D'un point

de vue syntactico-sémantique, ils sont caractérisés en termes de classes d'objets (<aliment>,<moyen de transport>, <outil>, <voie>, etc.) et de domaines (<aéronautique>, <médecine>, <sciences>, etc.) (Buvet et Mathieu-Colas, 1999). La macrostructure du dictionnaire est donc constituée de l'ensemble des noms correspondant à des arguments élémentaires. La microstructure comporte la vedette et des informations métalinguistiques relatives aux classes et aux domaines[9].

PRED-DIC : la structuration du dictionnaire des prédicats est plus complexe. Nous décrivons successivement la macrostructure puis la microstructructure.

Macrostructure

La nomenclature du dictionnaire PRÉD-DIC est constituée des racines prédicatives correspondant à autant d'emplois prédicatifs. Nous précisons successivement les notions de racine prédicative et d'emploi prédicatif.

Racine prédicative : la notion de racine prédicative rend compte du caractère polymorphe de certains prédicats : *Il l'aime*/*Il est amoureux d'elle*/*Il éprouve de l'amour pour elle*. La parenté morphologique entre le verbe, l'adjectif et le nom en position de prédicat et la stricte équivalence entre les trois énoncés permettent d'interpréter *aimer*, *amoureux* et *amour* comme trois formes différentes d'une même racine prédicative. Toutes les racines prédicatives ne donnent pas lieu à des énoncés équivalents : *Ceci gêne Luc*/*Ceci est gênant*/*Luc ressent de la gêne*.

Emploi prédicatif : un emploi prédicatif est défini conjointement par une racine prédicative, une classe sémantique et une interprétation donnée. Deux cas de figure sont à envisager selon que l'emploi prédicatif est autonome (*grognon*) ou polymorphique (*dédain*/*dédaigner*/*dédaigneux*).

Une racine prédicative polymorphique permet de rendre compte d'emplois prédicatifs non équivalents mais morphologiquement reliés et sémantiquement apparentés : *Luc déteste Max/Max est détesté/Max est détestable.* Les trois formes prédicatives ont une interprétation spécifique : *détester* s'interprète comme une 'propriété occasionnelle', *détesté* comme une 'propriété résultative' et *détestable* comme une 'propriété causale permanente'.

L'interprétation d'un prédicat est le produit de sa construction, de son trait et de son aspect inhérent. Si *détester*, *détesté* et *détestable* sont des prédicats qui partagent la même racine prédicative, en l'occurrence **détest-**, et appartiennent à la même classe sémantique <haine>, ils n'ont pas cependant la même interprétation.

1. celle du verbe résulte du fait qu'il a la construction **X0 V X1** avec **X0** ='humain' et **X1**='humain', le trait 'état' et l'aspect 'provisoire' ;
2. celle de l'adjectif participe tient au fait qu'il a la construction **X0 être Appé** avec **X0**= 'humain', le trait 'état' et l'aspect 'accompli' ;
3. celle de l'adjectif en *–able* s'explique parce qu'il a la construction **X0 être A** avec **X0**= 'humain', le trait 'état' et l'aspect 'permanent'.

Quels que soient les dictionnaires, les vedettes sont à l'intersection de leur macrostructure et de leur microstructure. Les vedettes de PRED-DIC sont des racines prédicatives correspondant à autant d'emplois prédicatifs. Autrement dit, une racine prédicative apparaît dans plus d'une entrée lorsqu'elle correspond à plus d'un emploi prédicatif. Par contre, un emploi prédicatif polymorphique est décrit sous la même entrée et il est spécifié dans l'article afférent quelles sont les différentes formes qu'il recouvre. Ainsi, *détestation* est dans le même article que *détester* du fait de l'équivalence des deux phrases suivantes :

Luc déteste Max

[9] Tous les noms du dictionnaire ne nécessitent pas d'être caractérisés par un domaine.

Luc a de la détestation pour Max

Nous présentons maintenant les informations métalinguistiques qui constituent le reste d'un article de PRED-DIC.

Microstructure

Les descripteurs associés à l'entrée d'un dictionnaire sont tous des propriétés linguistiques. Ils sont de trois ordres : les descripteurs de définition, les descripteurs de conditions et les descripteurs de validation

Les descripteurs de définition : il s'agit de la classe sémantique et de l'interprétation de l'emploi prédicatif correspondant à l'entrée. Elles constituent les deux informations qui sont associées à la racine prédicative lors de l'étiquetage syntactico-sémantique.

Les classes sémantiques qui caractérisent les prédicats sont en assez grand nombre (environ 2000 en l'état actuel des travaux du LLI). Par contre, les interprétations possibles sont limitées en nombre (une douzaine).

Les descripteurs de conditions : ce sont les différentes propriétés linguistiques qui permettent à l'étiqueteur sémantique de déterminer quels sont les emplois prédicatifs des racines prédicatives. Elles sont de cinq sortes.

Les propriétés morphologiques : ces propriétés sont au nombre de trois. Tout d'abord, elles indiquent les diverses formes simples possibles d'un même emploi prédicatif (le verbe *prendre* et le nom *prise* ou uniquement le verbe *prendre*). Elles font état de l'éventuel caractère complexe de l'entrée (*prendre ombrage*). Elles signalent aussi la défectivité (*prendre fin* ne s'emploie qu'à la troisième personne).

Les propriétés structurelles : elles font état du nombre des arguments et de leur mode de structuration. Le nombre de structures possibles dépend de la forme des prédicats. Par exemple, *prendre* est caractérisé par :

1. la construction **X0 V X1** en tant que synonyme de *commander* (*Luc prend une bière*) ;
2. la construction **X0 V X1 PREP2 X2** en tant que synonyme de *tenir* (*Luc prend Léa par la taille*) ;

Les propriétés distributionnelles : elles font état de la structure argumentale du prédicat en indiquant, d'une part, la nature syntaxique des arguments (groupe nominal, complétive, infinitive, etc.) et, d'autre part, la nature sémantique des prédicats (en termes de classes d'objets ou d'hyperclasses). Il est possible de la sorte d'établir que *prendre* a deux acceptions différentes selon que la position **X0** est occupée :

1. indifféremment par un groupe nominal ou une infinitive (*(Cette affaire Faire cela) prend du temps*) ;
2. seulement par un groupe nominal qui, de plus, correspond nécessairement à un humain (*Luc prend du temps*).

Les propriétés combinatoires : ces propriétés sont dissociées selon qu'elles ressortissent à la signification grammaticale (*brûler* dans *Luc brûle d'amour pour Léa*) ou bien à la signification lexicale (*intelligemment* dans *Luc a présenté le projet intelligemment*) (Blanco et Buvet, 2004).

Les propriétés paraphrastiques : il s'agit de reconstructions des phrases canoniques, typiquement le passif ou la forme pronominale. La construction standard *Luc prend Max au sérieux* donne la reconstruction du type passif *Max est pris au sérieux par Luc* et la reconstruction du type forme pronominale réfléchie *Luc se prend au sérieux*.

Elles sont symptomatiques de la polysémie des racines prédicatives dans la mesure où les reconstructions varient selon les emplois.

Les descripteurs de validation : ces descripteurs justifient les interprétations des racines constituant les entrées. Ils permettent de vérifier la cohérence de la définition proposée dans l'article. Si certains sont aussi des descripteurs de conditions (par exemple, les propriétés structurelles), les autres sont de propriétés sémantiques spécifiques qui ne participent pas directement à l'identification de l'emploi. Il s'agit du type (état, action événement) et l'aspect intrinsèque de l'emploi prédicatif (e.g. le ponctuel pour *gifler*).
La description paramétrée des différents emplois prédicatifs et des arguments donne à lieu à diverses informations explicites qu'il est possible de structurer sous un format informatisable.

4.2 Modélisation informatique

L'augmentation des performances des systèmes TAL est directement liée à celle des ressources linguistiques tant du point de vue qualitatif que du point de vue quantitatif. L'exploitation informatique de ces ressources, qu'elles soient de nature morphologique, syntaxique ou sémantique, est toujours problématique. Pour des raisons de réutilisabilité et de pérennité, il faut que les ressources respectent des normes reconnues par tous les acteurs du domaine (Francopoulo, 2006). La structuration informatique de ARGU-DIC et PRED-DIC utilise la pré-norme LMF (ISO, 2006). Celle-ci propose non pas une DTD XML toute faite mais plutôt un cadre dans lequel il possible de construire et documenter un grand nombre de ressources linguistiques décrites dans des formalismes très divers. Le risque de contraindre une représentation par le biais d'un DTD est l'abandon de celle-ci lorsqu'un modèle ne peut y trouver sa place. Nous nous appuyons également sur les recommandations de la TEI en ce qui concerne la représentations des dictionnaires « papier » et le codage des entêtes. La structuration métalinguistique de PRED-DIC est prise en charge au format XML comme suit :

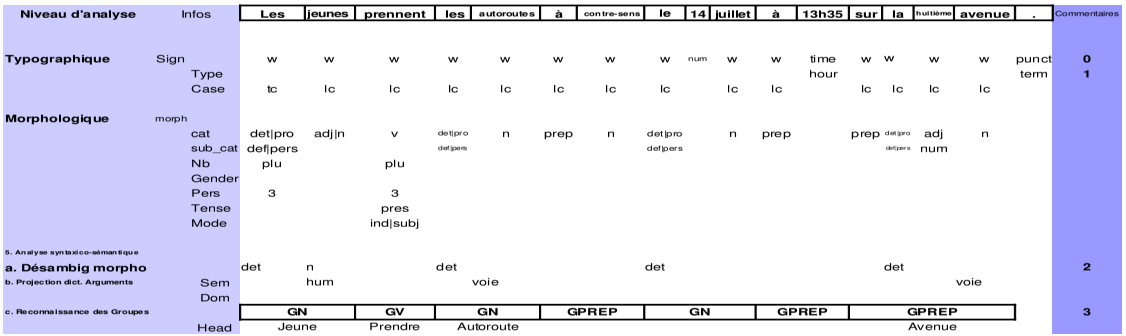
```
<entry id="prendre_1" class="capture" int="operation">
    <root>prendre</root>
    <example>les enfants prennent un chat
    </example>
    <definition>
        <item name="class" val="capture"/>
        <item name="int" val="operation"/>
    </definition>
    <morphProp>
        <item val="verb"/>
    </morphProp>
    <structprop>
        <item pred="verb" val="X0_V_X1"/>
    </structProp>
    <distProp>
        <struct type="syntax">
            <item arg="0" val="np"/>
            <item arg="1" val="np"/>
        </struct>
        <struct type="semantic">
            <item arg="0" val="hum"/>
            <item arg="1" val="animal"/>
        </struct>
    </distProp>
    <semanticProp>
        <item type="feature" val="action"/>
        <item type="aspect" val="perfective"/>
    </semanticProp>
    <appropriateProp></appropriateProp>
    <paraphrasticProp>
        <item pred="verb" val="passive"/>
    </paraphrasticProp>
</entry><classFrame>
        <item id="capture"/>
        <item id="hum"/>
        <item id="animal"/>
</classFrame>

<intFrame>
        <item id="operation"/>
</intFrame>

<interpretationFrame>
        <item id="operation"/>
</interpretationFrame>

<propertyFrame typepred="verb">
        <struct id="X0_V_X1">
                <item val="0"/>
                <item pred="verb"/>
                <item val="1"/>
        </struct>
        <struct id="X0_V_X1opt">
                <item num="0"/>
                <item pred="V"/>
                <item num="1" type="opt"/>
        </struct>
</propertyFrame>
```

5 Exemple de traitement[10]



Niveau d'analyse	Infos	Les	jeunes	prennent	les	autoroutes	à	contre-sens	le	14	juillet	à	13h35	sur	la	huitième	avenue	.	Commentaires
Typographique	Sign	w	w	w	w	w	w	w	w	num	w	w	time	w	w	w	w	punct	0
	Type												hour					term	1
	Case	tc	lc	lc	lc	lc	lc	lc	lc		lc	lc		lc	lc	lc	lc		
Morphologique	morph																		
	cat	det|pro	adj|n	v	det|pro	n	prep	n	det|pro		n	prep		prep	det|pro	adj	n		
	sub_cat	def|pers			def|pers				def|pers						def|pers	num			
	Nb	plu		plu															
	Gender																		
	Pers	3		3															
	Tense			pres															
	Mode			ind|subj															
5. Analyse syntaxico-sémantique																			
a. Désambig morpho		det	n		det				det						det				2
b. Projection dict. Arguments	Sem		hum			voie											voie		
	Dom																		
c. Reconnaissance des Groupes		GN		GV	GN		GPREP		GN			GPREP		GPREP					3
	Head	Jeune		Prendre	Autoroute									Avenue					

247

Commentaires (exemples de règles utilisées pour générer les représentations)

0 : les signes sont reconnus via des expressions régulières UNICODE.

Exemples : `w=(\p{L})+(-)(\p{L})+; punct=\p{P}`

1 : un certain nombre d'entités nommées à base numérique sont reconnues dès la reconnaissance typographique.

Exemple : `Hour=[0-9]{2}[h:][0-9]{2}`

2 : En français, la désambiguisation morpho-syntaxique porte notamment sur les déterminants/pronoms.

Exemple de règle de désambiguisation : `morph[1][@cat in 'det,pro'] and morph[@cat in 'n,adj'] and morph[3][@cat='v']=>det+n+v`

3 :la reconnaissance des groupes se fait sur la base de grammaires locales exprimées en XPATH :

Exemples : `w[1][@cat='det'] and w[2][@cat='n'] and w[3][@cat='v'] => GN[det n] V`

4 : Reconnaissance prédicative

Exemple : `GN[1][@sem='hum'] and GV[2][@head='prendre'] and GN[3][@sem='voie']`

[10] La phrase exemple est tirée du Monde en ligne, article du 19/07/2003

6 Conclusion

Nous avons présenté un prototype d'étiqueteur syntactico-sémantique intégré dans une plate-forme d'analyse linguistique. La particularité de cet étiqueteur est d'utiliser le modèle linguistique des classes d'objets. Celui-ci offre un moyen efficace de lier la syntaxe et la sémantique au sein d'une structure. Du point de vue informatique, nous avons décrit un système qui reprend les grandes étapes « classiques » en T.A.L., en nous appuyant sur des ressources linguistiques, et en mettant au point un langage d'expression de grammaires locales très proche d'une expression abstraite et quasi linguistique des phénomènes. L'implémentation proprement dite de ce modèle a été réalisé en respectant les normes (ou futures normes) et recommandations reconnues.

7 Bibliographie

BLANCO X et BUVET P.-A. (2004), « Verbes supports et significations grammaticales. Implications pour la traduction espagnol-français » in *Lingvisticae Investigationes* 27(2), John Benjamins B.V., Amsterdam

BUVET P.-A. (à paraître), « Détermination et figement au regard de la traduction », *META.*

BUVET P.-A. et MATHIEU-COLAS M. (1999), « Les champs *domaine* et *sous-domaine* dans les dictionnaires électroniques », *Cahiers de lexicologie*, 75, Didier Erudition, Paris, pp. 173-191.

SABAH G. (1996), Le sens dans les traitements automatiques des langues — le point après 50 ans de recherches, conférence invitée, journée ATALA (un demi-siècle de traitement automatique des langues : Paris.

GROSS G. (1995), « Une sémantique nouvelle pour la traduction automatique : les classes d'objets », in *La Tribune des Industries de la Langue et l'Information électronique*, 17-18-19, Paris.

GROSS G. et VIVES R. (1986), « Les constructions nominales et l'élaboration d'un lexique-grammaire », *Langue française*, 69, Larousse, Paris, pp. 5-27.

IDE, N., VERONIS, J., (1998). The state of the art. Computational Linguistics 24, 1–40, Introduction to the Special Issue on Word Sense Disambiguation

LE PESANT D. et M. MATHIEU-COLAS (1998), « Introduction aux classes d'objets » in *Langages* 131, Larousse, Paris.

MIHALCE R., EDMOND P. (Eds.), (2004). Proceedings of SENSEVAL-3 : Third International Workshop on Evaluating Word Sense Disambiguation Systems

PREISS J., STEVENSON M., (2004). *Word Sense Disambiguation,* Computer Speech & language,Volume 18, Issue 3

PIAO S.L., ARCHER D., MUDRAYA O., RAYSON P., GARSIDE R., MCENERY T., WILSON A. (2005) A Large Semantic Lexicon for Corpus Annotation. In proceedings of the Corpus Linguistics 2005 conference, July 14-17, Birmingham, UK. Proceedings from the Corpus Linguistics Conference Series on-line e-journal, Vol. 1, no. 1, ISSN 1747-9398.

VERONIS, J. (2004). « Quels dictionnaires pour l'étiquetage sémantique ? » Le français moderne, 72(1):27-38.

Session
Sémantique

Un analyseur hybride pour la détection et la correction des erreurs cachées sémantiques en langue arabe

Chiraz BEN OTHMANE ZRIBI, Hanène MEJRI , Mohamed BEN AHMED
Laboratoire de recherche RIADI, Université La Manouba
ENSI, La Manouba, Tunisie
Chiraz.benothmane@riadi.rnu.tn, Hanene.mejri@riadi.rnu.tn,
Mohamed.benahmed@riadi.rnu.tn

Résumé. Cet article s'intéresse au problème de la détection et de la correction des erreurs cachées sémantiques dans les textes arabes. Ce sont des erreurs orthographiques produisant des mots lexicalement valides mais invalides sémantiquement. Nous commençons par décrire le type d'erreur sémantique auquel nous nous intéressons. Nous exposons par la suite l'approche adoptée qui se base sur la combinaison de plusieurs méthodes, tout en décrivant chacune de ces méthodes. Puis, nous évoquons le contexte du travail qui nous a mené au choix de l'architecture multi-agent pour l'implémentation de notre système. Nous présentons et commentons vers la fin les résultats de l'évaluation dudit système.

Abstract. In this paper, we address the problem of detecting and correcting hidden semantic spelling errors in Arabic texts. Hidden semantic spelling errors are morphologically valid words causing invalid semantic irregularities. After the description of this type of errors, we propose and argue the combined method that we adopted in this work to realize a hybrid spell checker for detecting and correcting hidden spelling errors. Afterward, we present the context of this work and show the multi-agent architecture of our system. Finally, we expose and comment the obtained results.

Mots-clés : erreur cachée, erreur sémantique, détection, correction, système multi-agent, langue arabe.

Keywords: hidden error, semantic error, detection, correction, multi-agent system, Arabic language.

1 Introduction

Les erreurs cachées sont des erreurs orthographiques produisant des mots valides lexicalement et causant des dérèglements de haut niveau : syntaxique, sémantique, voire même pragmatique. Les erreurs cachées surviennent lorsqu'une ou plusieurs modifications sur un mot le transforme en un autre mot de la langue. Dans ce cas, l'erreur, est dans la plupart du temps, une graphie semblable au mot que l'utilisateur avait l'intention d'écrire.

*Le jardinier utilise le **gâteau** (râteau) pour bêcher la terre*

Dans cet exemple, le mot « *gâteau »* est introduit dans un contexte qui ne lui est pas approprié. Cette faute de frappe peut être corrigée en rétablissant le mot correct *« râteau ».*

Dans (Verberne, 2002) on lit que les statistiques réalisées pour la langue anglaise par (Eastman, Oakman, 1991) affirment que les erreurs cachées représentent 25% parmi toutes les erreurs orthographiques commises et contenues dans leur corpus de référence. (Mitton, 1987) cité par le même auteur, leur attribue une valeur plus grande à savoir : 40% parmi toutes les erreurs orthographiques étudiées. Ces deux valeurs assez importantes ont rendu l'étude de ce genre d'erreurs une nécessité en soi. Plusieurs recherches ont été entreprises dans le but de remédier à ce problème. Nous pouvons citer par exemple les recherches de Golding qui a étudié ce genre d'erreurs pour la langue anglaise. Il a ainsi proposé différentes méthodes comme la méthode de *Bayes* (Golding, 1995), la méthode des trigrammes des parties du discours (Golding, Schabes, 1996) et la méthode à base de réseaux neuronaux dite *Winnow* (Golding, Roth, 1999). Le chinois a été aussi traité avec les deux chercheurs (Xiaolong, Jianhua, 2001). Le suédois a également fait l'objet d'une recherche avec (Bigert, Knutsson, 2002).

En ce qui concerne la langue arabe, aucun autre travail n'a concerné le traitement des erreurs cachées malgré l'importance de l'entreprise d'une telle recherche. La langue arabe présente, en effet, des spécificités dont nous citons principalement : l'agglutination, l'ambiguïté grammaticale et la proximité lexicale. Toutes ces caractéristiques rendent le risque de commettre une erreur cachée plus important que pour les autres langues notamment latines.

Nous nous sommes donc intéressés à ce problème en construisant un système permettant à la fois de détecter et de corriger ce type d'erreurs pouvant survenir dans des textes arabes. Dans un premier temps ce système a concerné uniquement les anomalies syntaxiques (Ben Othmane et al., 2005). Nous l'avons amendé par la suite pour qu'il puisse traiter l'ensemble des anomalies (syntaxiques et sémantiques).

Dû à la complexité de ce travail, nous avons été amenés à émettre certaines hypothèses pour restreindre les champs de nos investigations. Nous avons considéré alors l'arabe non voyellé et ce pour une raison capitale. C'est que malgré l'importance des voyelles[1] dans la compréhension du discours arabe, elles n'apparaissent que très rarement dans les textes. Ainsi, à part quelques ouvrages poétiques ou littéraires didactiques, les écrits arabes sont généralement dépourvus de voyelles, et c'est le cas des textes fréquemment rencontrés dans les journaux, les revues, les romans, etc. Aussi, nous émettons l'hypothèse de l'existence d'une seule erreur par phrase et par mot. Cette erreur consisterait en une seule faute typographique du type : ajout d'un caractère, omission d'un caractère, substitution d'un caractère par un autre ou interversion de deux caractères adjacents. Des statistiques ont en effet montré que l'une (seulement) de ces opérations est à l'origine d'une erreur orthographique dans 90% des cas (Ben Hamadou, 1993).

Dans ce qui suit, nous décrivons dans la première section le type d'erreurs sémantiques auquel nous nous sommes intéressés et formant ce qu'on appelle des erreurs cachées sémantiques. Dans la deuxième section, nous présentons l'approche proposée pour la conception de notre système de détection–correction erreurs cachées sémantiques. Dans la troisième section de l'article, nous abordons le contexte de notre travail, ainsi, que l'architecture d'implémentation adoptée pour la réalisation de notre système. La quatrième et dernière section est consacrée, quant à elle, à la description des résultats de l'évaluation du système mis en place.

---

[1] Signes diacritiques ajoutées aux lettres arabes pour permettre leur lecture

2 Les erreurs cachées sémantiques

Nous entendons par « erreur cachée sémantique » tout mot ressemblant typographiquement à un caractère près au mot correct qu'il remplace mais invalide sémantiquement dans le contexte où il se trouve. Les dérèglements sémantiques causées par ce type d'erreurs peuvent être réparties en deux catégories: les *incompatibilités sémantiques* et les *incomplétudes sémantiques*. Quand l'erreur cause des contresens ou encore rend la phrase dépourvue de sens, nous parlons dans ce cas d'*incompatibilité sémantique*. Quand à l'incomplétude sémantique, elle concerne principalement l'oubli de mots, de syntagmes ou d'outils de coordination nécessaires à l'interprétation de la phrase.

Nous nous intéressons ici qu'aux anomalies mettant en cause le sens. Les erreurs d'incomplétude sont plus difficiles à déceler.

يعرضون عليه أموالا كبيرة (كثيرة)

Ils lui proposent de grandes (beaucoup) d'argent

Dans cette phrase erronée, l'adjectif "كبيرة" (grandes) est utilisé au lieu de l'adjectif "كثيرة" (beaucoup) et il se trouve dans un contexte inapproprié par la substitution de la lettre ب par la lettre ث.

3 Détection des anomalies sémantiques

Pour que la machine puisse traiter la sémantique des mots, elle doit disposer, par analogie à l'être humain, des connaissances à propos du sens des mots et des différents contextes dans lesquels ils apparaissent. Ces connaissances peuvent être obtenues à partir de plusieurs ressources informatiques telles que les dictionnaires sémantiques, les thésaurus, les réseaux sémantiques, les ontologies ou les corpus textuels.

Dans le cadre de ce travail, nous optons pour une solution basée sur l'apprentissage du sens des mots à partir des corpus textuels. Cette orientation repose sur un principe de la linguistique distributionnelle qui dit que : "*le sens d'un mot peut être défini statistiquement, à partir de l'ensemble des contextes (i.e., paragraphes, phrases, textes) dans lesquels ce mot apparaît*" (Landauer et al., 1998). Par exemple, le mot *avion* apparaît souvent conjointement avec des mots comme *décoller*, *aile*, *aéroport*, et rarement conjointement avec des mots comme *lion* ou *forêt*.

Pour détecter les erreurs cachées sémantiques, nous proposons une approche qui se base sur l'étude de la validité sémantique de chaque mot du texte à analyser dans son contexte et ceci par la combinaison de plusieurs méthodes permettant de représenter chaque mot en fonction du contexte proche et lointain dans lequel il apparaît et de comparer cette représentation aux représentations antérieures obtenues lors de l'apprentissage.

Nous faisons ainsi appel à quatre méthodes, de nature statistique ou mixte (linguistique et statistique), responsables chacune de vérifier la validité sémantique d'une phrase donnée. L'idée derrière cette combinaison est d'obtenir un analyseur d'erreurs cachées sémantiques capable de tirer profit des avantages de toutes les méthodes d'analyses sémantiques proposées. Ceci implique la construction de plusieurs systèmes de traitement d'erreurs cachées qui seront mis en confrontation quant à la sélection d'une erreur cachée sémantique dans une phrase. Cette confrontation est réalisée suite à l'application d'une procédure de vote qui prendra en considération tous les résultats issus de l'application des méthodes d'analyses sémantiques proposées et procèdera à un vote pour l'identification de l'erreur la plus probable garantissant ainsi une meilleure qualité d'analyse.

Pendant la phase d'apprentissage, sont récoltées à partir d'un corpus dit d'entraînement traité au préalable[2] toutes les connaissances nécessaires aux différentes méthodes proposées et formant leurs entrées. Ce corpus[3] comporte *30* textes de type économique, et compte environ *30 000* mots, *1827* phrases et *4029* lemmes. Les connaissances extraites se présentent sous forme de données linguistiques et statistiques et varient selon les besoins de chaque méthode d'analyse utilisée.

3.1 Méthode Coocurrence-Collocation

Cette méthode vérifie la validité contextuelle d'un mot en se basant sur sa probabilité contextuelle déduite du calcul des trois mesures suivantes :

- **Probabilité de cooccurrence :** Cette probabilité est calculée pour chaque mot m_i de la phrase à analyser pour une fenêtre de 10 mots[4]. Elle est exprimée par la formule de probabilité conditionnelle de Bayes suivante :

$$P(m_i|C) = P(m_i|c_k,.....c_{-1},c_1,.....c_k) = \frac{P(c_k,.....c_{-1},c_1,.....c_k|m_i)P(m_i)}{P(c_k,.....c_{-1},c_1,.....c_k)}$$

  Où m_i représente le mot à analyser, c_i les mots voisins du contexte proche et $P(m_i)$ la probabilité d'apparition du mot m_i dans le corpus d'apprentissage.

- **Coefficient de collocation :** Une collocation est une expression ayant une structure morphosyntaxique précise et une fréquence d'apparition importante dans le corpus d'apprentissage, exemple : شوارع المدينة (les rues de la ville). Pour calculer ce coefficient nous procédons d'abord à l'identification des collocations existantes dans une phrase en se basant sur une liste de collocations obtenue lors de la phase d'apprentissage. Pour se faire, nous avons utilisé et adopté une partie du système réalisé par (Mlayeh, 2004). Lorsque une collocation est identifiée dans une phrase, un coefficient collocationnel est attribué à chaque mot de cette expression. Ce coefficient n'est autre que la mesure de Kulczinsky, qui est un critère d'association permettant d'identifier le degré de corrélation de deux lemmes l_i et l_j calculée à l'aide de la formule suivante :

$$KUC = \frac{a}{2}\left(\frac{1}{a+b} + \frac{1}{a+c}\right)$$

  Où : a : le nombre d'occurrences du couple (l_i, l_j)
  b : le nombre d'occurrences des couples où l_i apparaît non suivi de l_j
  c : le nombre d'occurrences des couples où l_j est non précède dé l_i

  La valeur de ce coefficient varie entre 0 et 1 et il est égal à 0,5 quand l_i est toujours observé avec l_j. Une expression est considérée comme collocation si son coefficient de KUC est supérieur à 0,5.

- **Probabilité de répétition:** "*les mots ou plus précisément les lemmes des mots d'un texte ont tendance à se répéter dans le texte lui-même*". Cette hypothèse est déduite des comptages réalisés par (Ben Othmane, Ben Ahmed, 2003) sur un corpus textuel en langue arabe appartenant à un domaine particulier qui montrent qu'une forme

---

[2] Analysé morpho-syntaxiquement, découpé en phrases et en syntagmes nominaux et verbaux.

[3] Ces textes proviennent à l'origine du corpus de l'arabe contemporain collecté par Al-Sulaiti L. http://www.comp.leeds.ac.uk/eric/latifa/arabic-corpora.htm. Ils ont été choisis par ce qu'ils sont relatifs à un même domaine.

[4] La taille de la fenêtre est paramétrable et peut être facilement ajustée.

textuelle apparaît en moyenne 5,6 fois dans un même texte alors qu'un lemme apparaît en moyenne 6,3 fois et ce dans le même texte. Subséquemment, si le lemme d'un mot se répète très peu dans le texte, le mot en question peut correspondre à une erreur cachée. Cette probabilité concerne donc le taux d'apparition de chaque lemme des mots de la phrase, objet de vérification, dans le corpus de test. Ce taux est calculé par la formule suivante :

$$P(l_i) = \frac{\text{nombre d' occurences de } l_i}{\text{nombre total de lemmes}}$$

La combinaison de ces trois mesures en vue de l'obtention de la probabilité contextuelle $P(m_i)$ de chaque mot de la phrase se fait selon la formule linéaire suivante :

$$P(m_i) = \alpha * P(m_i|C) + \beta * KUC(m_i) + \delta * P(l_i)$$

Où $P(m_i|C)$ est la probabilité de cooccurrence du mot m_i, $KUC(m_i)$ est le coefficient collocationelle attribué à un mot m_i, $P(l_i)$ est la probabilité de répétition pour un lemme l_i du mot m_i. α, β, et δ sont des poids attribués aux différentes probabilités afin de mettre en évidence la contribution de chaque probabilité. Il est à noter que ces valeurs ne sont pas connues à l'avance et sont déterminées lors des expérimentations[5]. Toutefois, nous estimons que la valeur de α doit être plus importante que celles de β, et δ vu que le contexte voisin est plus déterminant pour le sens du mot à analyser que son contexte lointain.

Une fois les probabilités relatives à tous les mots de la phrase en question sont calculées, elles seront comparées à une valeur *seuil* déterminé lors des expérimentations. Le ou les vocables ayant une probabilité inférieure à ce *seuil* forment une liste d'erreurs cachées éventuelles.

3.2 Méthode Vecteur-Contexte

Cette méthode consiste à représenter chaque mot de la phrase par un vecteur en fonction du contexte dans lequel il apparaît. De ce fait, un vecteur mot Vm_i n'est autre qu'une représentation vectorielle de la probabilité de cooccurrence de ce mot avec chaque mot de la phrase. Considérons par exemple, la phrase suivante :

شرب الرجل كلبا(كأسا)

L'homme a bu un chien (un verre)

La matrice ci-dessus illustre la probabilité de cooccurrence de chaque mot m_i de la phrase avec les mots voisins de ce même contexte. Les colonnes de la matrice représentent les mots m_i et les lignes représentent les composantes du vecteur Vm_i. Ainsi, une cellule contient la probabilité de cooccurrence du mot m_i avec le mot m_j, calculée selon la formule suivante:

$$P(m_i \mid m_j) = \frac{\text{nombre de fois où } m_i \text{ et } m_j \text{ cooccurent}}{\text{nombre d' occurrence de } m_j}$$

$V_{\text{كلبا}}$

	كلبا	الرجل	شرب
شرب	0,3	0,6	
الرجل	0,1		0,6
كلبا		0,1	0,3

Tableau 1 : Matrice de cooccurrence des mots d'une phrase

[5] Pour nos expérimentations nous avons choisi: α=2, β=1 et δ =0,5.

Pour représenter le degré de corrélation de chaque mot m_i avec tous les autres mots m_j de la phrase, nous proposons de calculer la norme de chaque vecteur Vm_i exprimée comme suit :

$$\|Vm_i\| = \sqrt{\sum_{j=1}^{n} c_j^2}$$

Où c_j est la probabilité de cooccurrence du mot m_i avec le mot m_j de la phrase. Dans l'exemple précédent, les normes des vecteurs des mots شرب, الرجل, كلبا sont respectivement égales à 0,67 ; 0,6 et 0,31. Le mot ayant la norme la moins élevée est كلبا, est soupçonné d'une erreur cachée. D'une manière générale, nous évaluons la norme de chaque vecteur mot Vm_i à une valeur *seuil*. Le ou les mots ayant une norme inférieure au *seuil* sont ajoutés à la liste des mots suspectés.

3.3 Méthode Vecteur-Vocabulaire

Le vocabulaire (termes représentatifs) d'un texte ou d'un domaine en question est un élément caractéristique de ce dernier et un bon indicateur de la cohérence de ce texte. Nous pouvons, par conséquent et en adoptant le principe de représentation vectorielle précédemment cité, étudier la validité sémantique d'une phrase en représentant chaque mot lui appartenant par un vecteur en fonction de sa probabilité de cooccurrence avec le vocabulaire. Pour évaluer la proximité entre deux vecteurs, nous utilisons la métrique de distance angulaire exprimée comme suit :

$$Dist(Vm_i, Vm_j) = \arccos(Sim(Vm_i, Vm_j))$$

$$Sim(Vm_i, Vm_j) = \cos(Vm_i, Vm_j) = \frac{Vm_i \, Vm_j}{\|Vm_i\| * \|Vm_j\|} = \frac{\sum_{t=1}^{n} Vm_{it} \, Vm_{jt}}{\sqrt{\sum_{t=1}^{n} Vm_{it}^2} \sqrt{\sum_{t=1}^{n} Vm_{jt}^2}}$$

Le calcul de la distance angulaire se fait pour chaque vecteur mot m_i par rapport à tous les autres vecteurs mot m_j de la phrase. Le vecteur le plus éloigné du contexte correspond au mot qui apparaît le moins avec les mots du vocabulaire en corrélation avec le contexte courant. Pour sélectionner ce vecteur, la somme des distances angulaires de chaque vecteur mot m_i est calculée puis comparée à une valeur *seuil*. Le ou les mots qui correspondent à la somme des distances la plus élevée et supérieur au seuil sont soupçonnés d'erreurs cachées.

3.4 Méthode LSA

"LSA (Latent semantic Analysis : Analyse sémantique latente) est une méthode permettant l'acquisition des connaissances à partir de l'analyse entièrement automatique de grands corpus textuels" (Landauer et al., 1998). Plus précisément, cette méthode permet d'identifier la similarité sémantique entre deux mots, deux segments textuels ou la combinaison des deux même si ces mots ou segments textuels ne sont pas co-occurrents.

Le principe de la méthode LSA consiste à représenter les mots dits unités lexicales et les segments textuels (phrases, paragraphes, textes) dits unités textuelles par des vecteurs dans un espace vectoriel de dimensions réduites par rapport à l'espace d'origine et le mieux représentatif de ce dernier. L'espace d'origine est représenté par une matrice de cooccurrence initiale *X(m, n)* représentative du corpus d'apprentissage où les *m* lignes correspondent aux unités lexicales, et les *n* colonnes aux unités textuelles. Une cellule contient le nombre d'occurrences d'une unité lexicale dans une unité textuelle. Cette matrice est décomposée en produits de trois matrices *T(m,t)*, *S(t,t)* et *D(t,n)* grâce à une forme d'analyse factorielle appelée décomposition en valeurs singulières. La matrice *T* est une matrice orthogonale de *m*×*t* dimensions, *D* est une matrice orthogonale de *t*×*n* dimensions et *S* est une matrice

diagonale de $t \times t$ dimensions dite aussi matrice de valeurs singulières. Les valeurs de cette dernière représentent les dimensions de l'espace d'origine.

Dans notre cas, la matrice *X* a été construite durant la phase d'apprentissage. Les lignes correspondent aux lemmes dudit corpus, et ils sont au nombre de **4029**, les colonnes représentent les phrases dont le nombre est **1827**. La réduction des dimensions consiste à choisir parmi les *n* dimensions les *k* dimensions les plus pertinentes et les plus représentatives de l'espace d'origine à partir de la matrice diagonale *S* triée selon l'ordre de ses valeurs singulières. Ainsi, nous obtenons trois matrices *T (m,k)*, *S(k,k)* et *D(k,n)* de dimensions réduites (k=300 valeur choisie après plusieurs tests). Le produit scalaire de ces matrices génère la matrice *X'(m,n)* représentative de l'espace résultat.

La variante de la méthode *LSA* que nous proposons étudie la validité sémantique des mots d'une phrase donnée en comparant leurs vecteurs sémantiques extraits de la matrice de cooccurrence transformée et obtenue lors de la phase d'apprentissage. Pour mesurer la proximité sémantique entre les vecteurs issus de la matrice obtenue, nous utilisons, comme le cas de la méthode *Vecteur-Vocabulaire,* la métrique de distance angulaire. Ainsi, chaque vecteur sémantique Vm_i du mot m_i est comparé à tous les vecteurs Vm_j des mots m_j du contexte en fonction de la distance angulaire. La somme de ces distances est ensuite calculée pour chaque mot m_i et comparée à une valeur *seuil*. Si cette valeur est supérieure au *seuil*, le mot correspondant est soupçonné d'une erreur cachée.

3.5 Procédure de vote

Étant donné que notre système global de détection d'erreurs cachées se base sur l'hypothèse stipulant une erreur au plus par phrase et que les prétendues erreurs sont toujours classées par ordre de probabilité décroissante, nous avons choisi un vote de type *uninominal par classement* (les candidats sont triés et un seul parmi eux sera élu). Nous présentons dans ce qui suit le principe de la méthode que nous avons adoptée par notre procédure de vote.

1. Compter le nombre d'occurrences des différentes erreurs proposées par toutes les méthodes d'analyses sémantiques présentes dans chaque liste et se trouvant au premier rang.
2. Sélectionner les erreurs qui ont recueilli le plus grand nombre d'occurrences. Si une seule erreur obtient la majorité absolue du nombre d'occurrences, elle est élue comme étant l'erreur la plus probable dans la phrase. Sinon, on calcule une nouvelle valeur d'occurrences des erreurs retenues au rang suivant.
3. Ce processus se répète autant de fois jusqu'à ce qu'une seule erreur ayant la majorité absolue d'occurrences soit retenue.

Toutefois, la méthode de vote proposée peut conduire parfois à une situation de blocage où le nombre d'occurrence de deux ou plusieurs erreurs sélectionnées en premier rang reste toujours invariant. Dans ce cas, nous nous référons au *degré de confiance* attribué à chaque méthode afin de sélectionner, parmi la liste des erreurs retenues, celle détectée par la méthode du plus grand degré de confiance.

4 Correction des erreurs cachées sémantiques

Pour corriger les erreurs cachées, nous procédons à la génération de toutes les formes proches de la forme erronée, à un caractère d'édition près pour former ainsi une liste contenant les

candidats à la correction. Nous avons utilisé et adapté à cet effet un correcteur orthographique développé par (Ben Othmane, 1998).

Comme nous nous attendons à avoir un grand nombre de propositions, dû à la proximité lexicale de la langue arabe, nous avons pensé réduire cette liste. L'idée étant de substituer la forme erronée par chacune des formes proposées et former ainsi un ensemble de phrases candidates. Ces dernières seront soumises à notre détecteur d'erreur sémantique. Celles qui produisent des dérèglements dans la phrase seront éliminées et c'est le même sort que subissent leurs propositions respectives. La liste des propositions restantes est par la suite triée par ordre de pertinence et présentée à l'utilisateur.

5 Contexte de travail

Ce travail vient compléter nos recherches précédentes (Ben Othmane et al., 2005) qui ont concerné le problème d'erreurs cachées (syntaxiques et sémantiques) pouvant se produire dans un texte en langue arabe. Le système qui a été proposé pour le traitement de ces erreurs est à base d'agents. Ce système (SMA) se compose principalement d'un agent pour la correction et de deux groupes d'agents pour la détection : un groupe d'agents syntaxiques permettant de traiter les anomalies syntaxiques pouvant se produire dans une phrase donnée et un groupe d'agents sémantiques permettant de traiter les incohérences sémantiques. Seul l'agent correction et le groupe d'agents syntaxiques ont été bien étudiés et implémentés, nous venons donc compléter par notre travail la partie sémantique. La figure 1 illustre l'architecture globale du système de traitement des erreurs cachées.

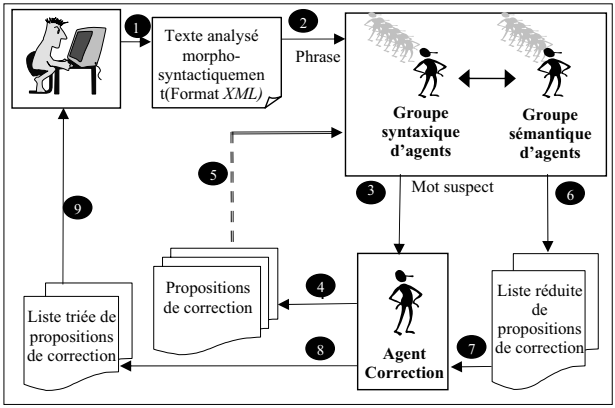


Figure 1 : Architecture du système global de détection et correction des erreurs cachées

Nous avons ainsi implémenté notre vérificateur sémantique sous forme d'un groupe d'agents sémantiques, où chaque méthode proposée est appliquée par un agent spécifique. En plus, un agent *Superviseur* du groupe est chargé de l'activation des différents sous agents sémantiques responsables d'analyser la phrase en cours et de détecter les incohérences sémantiques qu'elle peut renfermer. Les agents sémantiques travaillent en parallèle et communiquent leurs résultats à l'agent *Superviseur* qui joue en plus, dans ce cas, le rôle de décideur en sélectionnant l'erreur la plus probable parmi l'ensemble des listes d'erreurs détectées par les différents agents en appliquant la procédure de vote.

6 Expérimentations et résultats

Pour l'évaluation de notre système, nous avons choisi un texte de test de même type et appartenant au même domaine que le corpus d'apprentissage utilisé. Il compte 1 564 mots, 100 phrases dont 50 contiennent une erreur cachée.

La figure suivante illustre les performances de chaque agent, ainsi, que du système global de détection des erreurs cachées sémantiques en terme de précision.

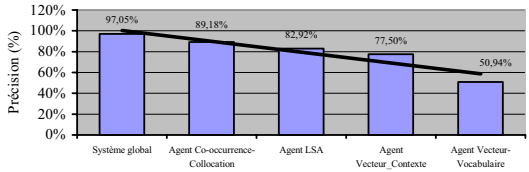


Figure 2 : Performances du système de détection des erreurs cachées sémantiques

Le taux de précision le plus élevé pour l'ensemble des agents sémantiques est celui de l'agent *Cooccurrence-Collocation* avec une valeur de **89,18%**. Cette performance s'explique par la complémentarité des phénomènes de cooccurrence, de collocation et de répétition. Contre toute attente, le taux fourni par l'agent *LSA* (**82,92%**) s'avère plus faible ; ceci est dû sans doute à la modestie de nos données d'apprentissage qui cause un taux élevé de sur-détection d'erreurs. Toutefois, la méthode *LSA* reste toujours prometteuse par rapport aux méthodes basées uniquement sur les cooccurrences des mots. En effet, le taux de précision de l'agent *Vecteur-Contexte*, est relativement faible (**77,5%)** et celui de l'agent Vecteur_Vocabulaire n'est pas bon (**50,94%**). L'amélioration des résultats de ces derniers nécessiterait à notre avis un grand corpus d'apprentissage, une stratégie d'extraction du vocabulaire du domaine plus fiable et une sélection fine et bien étudiée des textes formant le corpus d'apprentissage. Pour ce qui est du résultat de l'évaluation du système global, nous pouvons dire que le taux de précision qui est égal à **97,05%** est très satisfaisant**.** La performance du système de vote et son apport quant à la sélection de l'erreur la plus probable dans la phrase se confirment donc.

Quant à la phase de correction, elle a été testée à deux niveaux ; d'abord après l'obtention de toutes les propositions de correction, ensuite après la minimisation de la liste de ces propositions. Les résultats obtenus sont illustrés dans le tableau ci-après.

	Couverture	Précision	Ambiguïté	Proposition	Position
Initialement	100%	100%	100%	46,67	13,82
Minimisation	100%	80%	80%	5,98	3,43

Tableau 2 : Performance du système de correction des erreurs cachées sémantiques

Nous remarquons que notre méthode de minimisation de la liste des propositions a permis de réduire, considérablement (98%), le nombre moyen des propositions (46,67 à 5,98 propositions en moyenne). Cette diminution, bien qu'elle ait réduit l'ambiguïté de notre correcteur de 20%, ne s'est pas passée sans dégât. Elle s'est faite au dépend de la précision (diminution de 20%).

7 Conclusion

Notre système de détection d'erreurs cachées sémantiques a donné des résultats satisfaisants (taux de précision de **97,05%**) en dépit des contraintes et des restrictions liées à la taille ainsi qu'à la non diversité de nos données d'apprentissage. Nous signalons, aussi, l'apport de la démarche suivie pour la correction de la forme erronée qui a permis de minimiser la liste des propositions de correction de **98%** et d'avancer la forme correcte aux premiers rangs. Cependant, nous estimons que les résultats obtenus peuvent être encore améliorés d'abord par l'utilisation d'un bon corpus d'apprentissage de nature plus varié et de taille plus importante. D'autres perspectives proches sont également en vue, nous pensons effectivement intégrer les deux groupes d'agents syntaxiques et sémantiques ensemble afin de former le système global de traitement des erreurs cachées en langue arabe.

Références

BEN HAMADOU A. (1993). Vérification et correction automatique par analyse affixale des textes écrits en langue naturelle : le cas de l'arabe non voyellé. Thèse d'état en informatique, Faculté des Sciences de Tunis.

BEN OTHMANE Z. C. (1998). De la synthèse lexicographique à la détection et la correction des graphie fautives arabes. Thèse de doctorat, Université de Paris XI, Orsay.

BEN OTHMANE Z. C., BEN AHMED M. (2003). Le contexte au service des graphies fautives arabes. TALN'03, Batz-sur-Mer.

BEN OTHMANE Z. C., BEN FRAJ F., BEN AHMED M. (2005). Un système multi-agent pour le traitement des erreurs cachées en langue arabe. Actes de la 12ème Conférence sur le Traitement Automatique des langues naturelles TALN'05, Dourdan, vol. 1, p. 143-153.

BIGERT J., KNUTSSON O. (2002). Robust Error Detection : A Hybrid Approach Combining Unsupervised Error Detection and Linguistic Knowledge. In Proceedings of Robust Methods in Analysis of Natural Language Data (ROMAND'02), Frascati, Italie.

GOLDING A. (1995). A Bayesian hybrid method for context-sensitive spelling correction. In Proceedings of the third Workshop On Very Large Corpora, Cambridge, Massachuses, USA, (1995), 39-53.

GOLDING A., SCHABES Y. (1996). Combining trigram based and feature based methods for context sensitive spelling correction. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, 71-78.

GOLDING A., ROTH D. (1999). A winnow-based approach to context-sensitive spelling correction. Machine Learning, 34(1-3), 107-130.

LANDAUER T.K., Foltz P.W., Laham D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, Vol. 25, 259-284.

MLAYEH I. (2004). Extraction de collocations à partir de corpus textuels en langue arabe. Mémoire de mastère, Ecole nationale des sciences informatiques, Université de la Manouba.

VERBERNE S. (2002). Context sensitive spell checking based on word trigram probabilities. Master thesis Taal, Spraak & Informatica, University of Nijmegen.

XIAOLONG W., JIANHUA L. (2001). Combine trigram and automatic weight distribution in Chinese spelling error correction. Journal of computer Science and Technology, Volume 17 Issue 6, Province, China.

Résolution de la référence dans des dialogues homme-machine : évaluation sur corpus de deux approches symbolique et probabiliste

Alexandre DENIS[1], Frédéric BÉCHET[2], Matthieu QUIGNARD[1]
[1] UMR 7503 LORIA/CNRS – Campus scientifique
56 506 Vandoeuvre-lès-Nancy Cedex
[2] LIA – 339, chemin des Meinajaries, BP 1228,
84 911 Avignon Cedex 9
{alexandre.denis,matthieu.quignard}@loria.fr,
frederic.bechet@univ-avignon.fr

Résumé. Cet article décrit deux approches, l'une numérique, l'autre symbolique, traitant le problème de la résolution de la référence dans un cadre de dialogue homme-machine. L'analyse des résultats obtenus sur le corpus MEDIA montre la complémentarité des deux systèmes développés : robustesse aux erreurs et hypothèses multiples pour l'approche numérique ; modélisation de phénomènes complexes et interprétation complète pour l'approche symbolique.

Abstract. This paper presents two approaches, one symbolic, the other one probabilistic, for processing reference resolution in the framework of human-machine spoken dialogues. The results obtained by both systems on the French MEDIA corpus points out the complementarity of the two approaches : robustness and multiple hypotheses generation for the probabilistic one ; global interpretation and modeling of complex phenomenon for the symbolic one.

Mots-clés : dialogue homme-machine, résolution de la référence, évaluation, compréhension dans le dialogue.

Keywords: human-machine dialogue, reference resolution, dialogue understanding, evaluation.

1 Introduction

Le projet MEDIA de l'action TECHNOLANGUE propose une méthodologie d'évaluation des systèmes de compréhension de la parole dans un cadre de dialogues homme-machine. Le paradigme est inspiré du projet PEACE (Maynard & Devillers, 2000). Il définit une représentation sémantique commune vers laquelle chaque système devra convertir sa propre représentation.

A partir du corpus collecté (Bonneau-Maynard *et al.*, 2005), le projet MEDIA se divise en deux campagnes d'évaluation : MEDIA-HC (hors contexte) et MEDIA-EC (en contexte). Dans la première campagne les systèmes sont évalués sur leur capacité à produire la forme sémantique désirée sans prendre en compte le contexte du dialogue, chaque énoncé étant considéré comme

indépendant (Bonneau-Maynard *et al.*, 2006). La deuxième campagne évalue la compréhension d'un énoncé dans le contexte du dialogue. En particulier deux aspects sont évalués :

– spécification du sens en contexte des entités détectées lors de la phase MEDIA-HC ;
– résolution des références vers des entités introduites dans des tours précédents du dialogue.

Deux systèmes ont participé à cette évaluation en-contexte, un système réalisé au LIA et un autre au LORIA, chacun implémentant des méthodes différentes : une approche numérique pour le LIA, une approche symbolique pour le LORIA. Cet article décrit ces deux systèmes, l'outil d'évaluation développé pour cette campagne par le LORIA et les premiers résultats obtenus.

2 Problématique et manuel d'annotation

La résolution de la référence est le processus cognitif qui met en relation une expression (linguistique ou non) et la représentation mentale que cette expression désigne (voir pour une définition proche (Reboul & Moeschler, 1994)). Une représentation mentale est définie comme l'aggrégation de données hétérogènes (perceptives, mémorielles, logiques, etc.) sur une entité (Reboul & Gaiffe, 1999). Dans cette campagne, nous nous focalisons sur la représentation des référents acquise par une relation intra-linguistique de *coréférence*, définie entre deux expressions référentielles si ces dernières désignent le même référent (van Deemter & Kibble, 2000).

L'évaluation de la référence au sein du consortium devait répondre à deux objectifs principaux : elle devait être accessible à tous les systèmes participants et elle devait être compatible avec le paradigme d'évaluation hors-contexte et en particulier la représentation sémantique utilisée (Bonneau-Maynard *et al.*, 2005).

La plupart des approches évaluent les relations entre expressions référentielles (Popescu-Belis *et al.*, 2004), et s'appuient sur des formats d'annotation qui se concentrent sur les relations, à l'instar du format des campagnes MUC-6 et MUC-7 basé sur les coréférences (Chinchor & Hirschmann, 1997; van Deemter & Kibble, 2000). Dans cette lignée, le Reference Annotation Framework, RAF (Salmon-Alt & Romary, 2004) définit des catégories de données pour annoter les expressions référentielles (les *markables*) et les relations de différentes natures qu'elles entretiennent (les *referentials links*).

Cependant, comme tous les systèmes ne pouvaient produire ces relations mais étaient tous capables de fournir une représentation sémantique des référents (Bonneau-Maynard *et al.*, 2006), il a été décidé d'évaluer non pas les relations mais les descriptions des référents. L'annotation en contexte est alors vue comme une extension de l'annotation hors-contexte. Nous nous sommes toutefois inspirés des catégories pertinentes de liens référentiels définis dans RAF (identité, codomanialité, partie-tout) pour annoter les formes d'expressions référentielles.

Annotation hors-contexte de la référence. L'annotation hors-contexte consiste à annoter chaque segment signifiant d'un énoncé par un trait sémantique du type ⟨*mode*, *attribut*, *valeur*⟩ dont les contraintes ont été définies collectivement lors de l'évaluation hors-contexte [1] (Bonneau-Maynard *et al.*, 2005). En ce qui concerne la référence, cette annotation s'est limitée à annoter la présence d'une expression référentielle grâce à un trait d'attribut *lienRef* raffiné

[1] le mode décrit la polarité du trait (positive, négative, interrogative, optionnelle), l'attribut représente la catégorie sémantique mentionnée, et la valeur son instanciation

par la catégorie de l'expression. Les différents raffinements retenus (appelés *spécifieurs*) sont proches des catégories de RAF (voir tableau 1), à l'exception de partie-tout que nous n'avons pas annoté faute de consensus. A la différence des *markables* de RAF, seuls les déterminants des groupes nominaux sont associés au *lienRef* afin de ne pas interférer avec le reste du groupe nominal déjà annoté en sémantique.

TAB. 1 – Types de spécifieurs

Spécifieur	Signification	Expressions référentielles
coRef	coréférence : l'expression référentielle désigne son référent par référence directe	pronoms, définis, démonstratifs
elsEns	élément-ensemble : l'expression référentielle désigne son référent en vertu de propriétés sémantiques ou indexicales qui l'opposent à d'autres entités dans un ensemble	ordinaux, superlatifs, relatives, certains pronoms démonstratifs
coDom	co-domaine : l'expression référentielle désigne son référent grâce à un marqueur linguistique d'altérité	altérités

Afin de réduire le coût d'annotation, seules les expressions référentielles dont la résolution dépasse le cadre de l'énoncé ont été annotées. Cela exclut dès lors les référents dont l'antécédent a été introduit dans le même énoncé, les entités nommées et les indéfinis[2]. En revanche, les articles définis ont été annotés systématiquement, du moins pour les entités relevant de la tâche.

Annotation en contexte de la référence. Une référence est représentée comme un ensemble de référents, chacun décrit par un ensemble de traits sémantiques. On adjoint un champ *reference* à tous les traits *lienRef*. On notera par exemple {(t1,t2), (t3)} une expression référentielle qui fait référence à deux entités, l'une décrite par deux traits et l'autre décrite par un seul.

Exemple d'annotation hors et en contexte :

je veux / une / chambre double / et / une / chambre simple
+/nombre-chambre : 1
+/chambre-type : double
+/nombre-chambre : 1
+/chambre-type : simple
est-ce que / **ces** / chambres / ont la douche ?
+/lienRef-coRef : pluriel **reference = {**
(+/nombre-chambre :1, +/chambre-type :double),
(+/nombre-chambre :1, +/chambre-type :simple)}
+/objet : chambre
?/chambre-equipement : douche

Afin de raffiner l'expression du formalisme on autorise le *lienRef* a être respécifié dans certains cas. Par exemple, pour représenter l'ambiguïté, en l'absence d'un niveau supplémentaire requis, on approxime l'alternative comme un ensemble de référents en spécifiant le *lienRef* par *ambigu*. Ou encore pour distinguer les référents exclus des référents inclus (en particulier pour l'indéfini avec altérité, voir note 2) on spécifie le *lienRef* par *exclusion* ou *inclusion*.

Règles d'annotation. Nous avons collectivement défini des règles d'annotation pour réduire le risque de désaccord entre annotateurs et s'adapter aux spécificités de l'annotation hors-contexte [3]. Tout

[2] À l'exception de l'indéfini avec altérité (par ex. "une autre chambre") que nous avons jugé pertinent d'évaluer. Dans ce cas ce n'est pas le référent qui est annoté (il est indéfini) mais l'entité *exclue*.

[3] ces règles sont disponibles sur `http://www.loria.fr/~denis/media.html`, l'accord inter-annotateur est donné dans le tableau 2

d'abord la portée de la description d'un référent est constituée uniquement des traits sémantiques présents dans le dialogue antérieur en excluant l'énoncé courant et les énoncés simultanés en cas de chevauchements. Ensuite la couverture de la description est un compromis entre une annotation maximale qui aurait été trop coûteuse à effectuer et une annotation minimale restreinte aux traits discriminants peu intéressante à évaluer lorsqu'il n'y a qu'un seul référent. Nous avons alors distingué entités nommées et non-nommées :

– les entités nommées ou assimilées (hôtel nommé, dates, prix, villes, etc.) ne sont décrites que par un ensemble très restreint de traits comme leur nom ou leur valeur ;
– les entités non-nommées (hôtel non nommé et chambre) sont, elles seules, annotées avec la totalité de la description possible, y compris les traits d'autres référents.

Enfin on contraint la description des référents à être en forme normale, c'est-à-dire principalement non-redondante et non-contradictoire, de manière à homogénéiser les descriptions répétées (comme dans les anaphores fidèles) ou révisées.

3 L'approche symbolique du LORIA

Le système de résolution de la référence développé au LORIA est fondé sur la théorie des domaines de référence (Corblin, 1987; Reboul *et al.*, 1997; Salmon-Alt, 2001). Il postule que la désignation des référents passe par l'identification préalable d'un domaine dans lequel l'expression référentielle isole le référent. Cette vision de la référence se rapproche de la théorie des espaces focaux de Sidner (Grosz & Sidner, 1986). Elle a pour but d'unifier le traitement des différentes modalités ou types de désignation modélisés de manière hétérogène par d'autres systèmes. En particulier les ordinaux et les expressions d'altérité ne nécessitent pas de traitement *ad hoc* et s'intègrent élégamment dans la théorie. L'évaluation MEDIA fut pour nous l'occasion de tester le modèle sur l'anaphore bien qu'il fût à l'origine créé pour la référence multimodale.

Structuration en domaines de référence. Un domaine de référence est constitué d'un support, un ensemble d'objets défini intensionnellement ou extensionnellement, et d'un ensemble de critères de différenciation qui en discriminent les éléments. Chaque désignation active le domaine dans lequel on extrait le référent en le focalisant, le préférant ainsi pour les désignations ultérieures. On représente ici un domaine comme un couple (S, C), où S est l'ensemble support et C un ensemble de critères (formalisés par des relations d'équivalence). Chaque critère sera noté $c : F$, où c est la relation et F un élément focalisé de la partition opérée par c. Dans l'exemple suivant un seul critère (l'index) sera utilisé pour discriminer les référents.

S : *je vous propose* l'hôtel ibis *et* l'hôtel lafayette	`H*=( {h1, h2}, {index :{}} )`
U : *est-ce que* le premier hôtel *accepte les animaux*	`H*=( {h1, h2}, {index :{h1}} )`
S : *non* l'hôtel ibis *n'accepte pas les animaux*	`H*=( {h1, h2}, {index :{h1}} )`
U : *ok, je prends* l'autre *alors*	`H*=( {h1, h2}, {index :{h2}} )`

L'expression ordinale "le premier hôtel" réfère à h1 en vertu de son index, ce dernier reçoit donc la focalisation pour le critère "index". L'expression d'altérité recherche un domaine présentant une partition focalisée dans laquelle extraire l'autre élément, ce qui a pour effet de focaliser h2 dans la partition "index" (Denis *et al.*, 2006b).

Projection dans le formalisme. La projection consiste à construire la représentation MEDIA d'un référent. Etant donné que les domaines de référence ne conservent que le point de vue courant sur les référents et non pas les expressions référentielles et le contexte de leur emploi, il était nécessaire de

combiner le modèle avec une représentation lexicale et sémantique des référents. Parallèlement à la structuration domaniale de l'espace référentiel, nous avons conservé la structure sémantique des énoncés à laquelle nous avons ajouté des relations référentielles. Cette dernière s'appuie sur le MultiModal Interface Language, MMIL (Landragin & Romary, 2004) la représentation sémantique utilisée lors de la phase hors-contexte qui permet la représentation d'informations de natures lexicale, syntaxique ou sémantique. Ces liens référentiels nous permettent de parcourir les chaînes de coréférence afin d'annoter les descriptions des référents en fonction de leur représentation sémantique à différents instants du dialogue.

L'algorithme est le suivant :

1. interprétation de l'énoncé et production d'une forme sémantique (Denis *et al.*, 2006a) ;
2. résolution de chaque référent de l'énoncé : identification d'un domaine compatible grâce à la forme sémantique puis extraction et focalisation du référent (Salmon-Alt, 2001) ;
3. mise à jour de l'historique sémantique : création des relations d'anaphores (coréférence, et anaphore associative) entre les instances des référents ;
4. projection par parcours des chaînes de coréférence : aggrégation des informations sémantiques hors-contexte relatives aux référents dans un voisinage prédéfini correspondant au type d'entités (présence d'une relation sémantique, co-occurrence dans le composant sémantique, co-occurrence dans l'énoncé, présence dans le dialogue antérieur).
   Dans une des conditions du test (*avecHC*, cf. §5), nous intégrons à ce stade les annotations HC fournies par le protocole. Nous n'en tirons pas partie pour identifier les référents.

4 L'approche probabiliste du LIA

Le système d'interprétation du LIA est composé de deux niveaux. Le premier niveau effectue le décodage conceptuel en calculant, à partir d'un message vocal, une liste des n-meilleures séquences de concepts (où chaque concept est représenté par trois champs qui sont la chaîne de mots supports, l'attribut sémantique et la valeur). L'attribution des spécifieurs, du mode et la résolution des références sont pris en compte par le second niveau prenant en entrée la liste des n-meilleures séquences de concepts obtenues à partir du message vocal. Ce système (Servan & Bechet, 2006) est fondé sur une approche probabiliste identique à celle utilisée en Reconnaissance Automatique de la Parole (RAP). Ce système a été développé pour traiter directement des messages vocaux, cependant les résultats présentés dans cette étude sont obtenus sur les transcriptions manuelles du corpus, comme pour toute la campagne MEDIA.

La compréhension en contexte est ici effectuée selon l'approche suivante : la spécification du sens est vue comme une tâche d'étiquetage pouvant être traitée grâce à des techniques d'étiquetage probabilistes. La résolution des références est faite par un certain nombre d'heuristiques décrivant tous les rattachements possibles pouvant être faits entre la liste des n-meilleures interprétations et l'historique du dialogue.

Spécification du sens en contexte. Les séquences de concepts produites à partir du graphe de concepts dans la phase de décodage conceptuel ne contiennent aucun spécifieur de sens, hors ou en contexte. Ces spécifieurs sont attribués par un étiqueteur probabiliste basé sur les Champs Conditionnels Aléatoires (ou *Conditional Random Fields*, CRF). Les CRF (Lafferty *et al.*, 2001) ont été utilisés avec succès dans de nombreuses tâches d'étiquetage telles que l'étiquetage morphosyntaxique 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. Cela est particulièrement intéressant pour l'étiquetage des spécifieurs dans la mesure où la spécification du sens

d'un concept peut se faire avec des éléments situés avant ou après le concept dans l'énoncé courant, ou dans les énoncés précédents. Les liens référentiels sont étiquetés dans cette phase par rapport au type de l'objet référencé et au type de référence.

Le corpus d'apprentissage des CRF est obtenu à partir des corpus MEDIA. Chaque dialogue constitue une séquence où les observations sont les concepts marqués dans la référence et les étiquettes sont soit les spécifieurs attribués aux concepts ; soit le type du ou des objets référencés pour les liens référentiels, ainsi que le type du lien ; soit le symbole NULL si un concept n'a ni spécifieur ni lien référentiel. Lors du traitement d'un message, chaque chaîne de concepts produite par le décodeur mot/concept est traitée par l'étiqueteur CRF et la description des concepts est enrichie avec les étiquettes attribuées. L'étiqueteur développé utilise l'outil `CRF++`[4].

Résolution de la référence. A la suite de la phase précédente les concepts liens référentiels sont étiquetés avec les trois informations suivantes : le type de ou des objets pointés (*chambre*, *hôtel*, *réservation* ou une combinaison de ces valeurs) ; le type du lien référentiel (*ambigu*, *exclusion*, *inclusion*) ; et enfin le nombre (*singulier* ou *pluriel*).

La résolution des références s'effectue alors selon l'algorithme suivant : tous les concepts situés dans l'historique du dialogue (limité aux n énoncés précédents) ayant une étiquette *spécifieur* similaire au type d'objet pointé par le lien référentiel sont associés à ce lien.

Chaque objet est caractérisé par un certain nombre de traits (par exemple la ville, la marque, le nom ou les services associés à un hôtel). L'algorithme d'association fait pointer le lien référentiel vers tous les concepts représentant ces traits. Lorsque tous les traits sont identifiés, l'algorithme s'arrête s'il s'agit d'un lien référence singulier. Sinon un nouvel objet est créé et le processus se poursuit. Aucun contrôle n'est effectué sur le nombre d'objets désignés. Le but ici est de maximiser les mesures de rappel sur les références pour proposer au module de décision (analyseur sémantique, gestionnaire de dialogue) le plus d'associations possibles, chacune avec un score donné par les différents modèles utilisés lors du décodage.

5 Méthodologie

Dans le cadre de la campagne MEDIA un corpus de 1 250 dialogues a été constitué selon un protocole de type *Magicien d'Oz* : 250 locuteurs ont effectués chacun 5 scénarios de réservation d'hôtel avec un système de dialogue simulé par un opérateur humain. Pour la campagne MEDIA-EC le corpus d'apprentissage disponible contient 814 dialogues, 11 800 énoncés utilisateurs et 38 800 segments sémantiques dont 2 294 liens référentiels. Le corpus de test MEDIA-EC contient 174 dialogues pour 2 650 énoncés utilisateurs et 7 780 segments dont 910 liens référentiels.

Deux conditions de test sont organisées : la première (*sansHC*) consiste à n'utiliser que les dialogues transcrits, les erreurs de l'étiquetage hors-contexte se cumulant à celles de l'étiquetage en-contexte. La deuxième condition (*avecHC*) consiste à utiliser les dialogues avec leur annotation hors contexte.

Méthode d'évaluation. L'évaluation de la résolution de la référence s'effectue par comparaison des traits sémantiques proposés pour chaque référent. Nous observons que pour décrire un référent, il faut préalablement l'avoir identifié. De même, pour l'identifier, il faut préalablement avoir repéré l'expression référentielle. Comme ces tâches impliquent potentiellement des capacités différentes, nous avons jugé intéressant d'évaluer la résolution de la référence selon quatre niveaux, chacun donnant lieu à des scores de rappel, précision et f-mesure :

[4] Téléchargeable à `http://chasen.org/~taku/software/CRF++`

IER Capacité à repérer (ou identifier) les expressions référentielles. Il s'agit d'une capacité hors contexte, qu'on évalue tout de même car l'évaluation de la référence en dépend.

DER Capacité à décrire les expressions référentielles identifiées, c'est-à-dire, à fournir les bons spécifieurs (*coRef*, *coDom*, *elsEns*, mais aussi *inclusion*, *exclusion* et *ambigu*). Cette capacité est évaluée sur la base des expressions correctement repérées en IER.

IREF Capacité à identifier les référents, c'est-à-dire à fournir pour chaque référent suffisamment de traits corrects pour qu'il soit couplable avec un référent à trouver. Comme la précédente, cette capacité n'est évaluée que sur les expressions référentielles correctement repérées en IER.

DREF Capacité à décrire *in extenso* les référents. Cette évaluation n'est calculée que sur les référents corrects en IREF.

En procédant ainsi par niveau, nous pouvons mieux apprécier les différentes capacités qu'implique la résolution de la référence. Nous notons qu'il est possible d'avoir un score global de rappel (resp. de précision) en DREF, c'est-à-dire le nombre de traits corrects fournis par un système rapporté au nombre de traits fournis par l'annotation manuelle (resp. fournis par le système), en multipliant les scores de rappel (resp. de précision) obtenus en IER, IREF et DREF.

Pour chaque niveau, nous avons développé les algorithmes suivants :

IER Le but est d'aligner les traits *lienRef* sans s'appuyer ni sur leurs spécifieurs (évalués en DER), ni sur leur contenu référentiel (évalué en IREF et DREF). Or, si l'on retire ces annotations, le trait *lienRef* comporte trop peu d'information pour pouvoir effectuer l'alignement sans risque dans le cas d'une omission ou d'une addition de *lienRef*. Nous effectuons donc un alignement des *lienRef* sur la base des autres traits sémantiques qui sont dans l'intervalle. Nous avons adapté l'algorithme de Levenshtein pour que le gain d'un appariement de *lienRef* soit proportionnel à la valeur d'appariement des traits (non référentiels) qui se trouvent entre un *lienRef* et le suivant[5].

DER Pour chaque couple de *lienRef* appariés selon le procédé ci-dessus, on calcule le nombre d'erreurs qui apparaissent lorsqu'on rajoute les modes et surtout les spécifieurs.

IREF Pour chaque couple de *lienRef* appariés en IER, on effectue un couplage maximal de poids maximal entre référents hypothèse et référents de l'annotation manuelle. La matrice de couplage donne un poids proportionnel aux nombres de traits partagés[6].

DREF Pour chaque couple de référents formé en IREF, on effectue un couplage maximal entre traits. En effet, il n'y a pas d'ordre prescrit pour décrire les référents.

Evaluation de la qualité de l'annotation manuelle. L'annotation EC a fait l'objet de contrôles à trois reprises. A chaque fois, une double annotation a été effectuée sur une dizaine de dialogues puis évaluée à l'aide de l'outil de mesure présenté plus haut. Cette évaluation repose donc sur un échantillon ne représentant que 4% du corpus d'apprentissage (31 dialogues sur 814).

IAG	Dialogues	IER	DER	IREF	DREF
1	10	24 (24)	24 (24)	23 (25)	23 (23)
2	10	29 (29)	27 (29)	32 (33)	72 (108)
3	11	27 (27)	25 (27)	34 (36)	135 (150)
Total	31	80 (80)	76 (80)	89 (94)	230 (281)
		100%	95%	95%	82%

TAB. 2 – Accord inter-annotateurs

Les résultats sont présentés dans le tableau 2. L'accord inter-annotateurs est globalement très bon, surtout dans les niveaux supérieurs (IER, DER et IREF). Le score en DREF, plus faible que les autres, traduit la difficulté de fournir unanimement une description complète des référents.

[5]Ce procédé évite aux systèmes de produire une double annotation (HC et EC), l'IER devant être *a priori* calculée sur les *lienRef* corrects en HC et non sur l'annotation EC.

[6]Il suffit donc qu'un référent ait un trait correct pour être candidat. Pour être plus précis, il faudrait ne conserver que les traits permettant de discriminer un référent parmi les autres référents proposés.

6 Résultats

TAB. 3 – Résultats en précision et rappel de la tâche de résolution des références pour les systèmes du LIA et du LORIA. L'accord est une f-mesure des scores de rappel d'un système par rapport à l'autre.

	LIA		LORIA		Accord
sansHC	*prec*	*rappel*	*prec*	*rappel*	*f-mes*
DER	71.4	71.4	50.9	50.9	51.6
IREF	74.1	61.9	65.2	44.3	49.8
DREF	67.3	55.2	68.9	48.3	53.3

	LIA		LORIA		Accord
avecHC	*prec*	*rappel*	*prec*	*rappel*	*f-mes*
DER	86.5	86.5	86.5	86.5	86.5
IREF	77.1	73.8	62.4	40.8	44.0
DREF	74.1	64.0	76.5	43.3	51.6

Le tableau 3 présente les résultats des systèmes du LIA et du LORIA pour les deux conditions du test : *sansHC* et *avecHC*. En DER, le LORIA a un score médiocre dans la phase *sansHC*, qui s'améliore considérablement par la connaissance des traits HC. Pour l'identification des référents (IREF), le système symbolique pèche considérablement en rappel, avec un score comparable dans les deux phases *sansHC* et *avecHC*. Cela s'explique par le fait que l'annotation HC n'est pas utilisée pour la résolution de la référence proprement dite et n'intervient que pour améliorer la description des référents lorsque ceux-ci ont été identifiés. À l'inverse le système statistique tire bien meilleure partie des informations hors-contexte et parvient à augmenter ses scores de rappel d'une dizaine de points sur les trois catégories.

Pour investiguer plus profondément la complémentarité des systèmes, nous les avons évalués l'un par rapport à l'autre. Les résultats sont donnés dans la dernière colonne de chaque tableau. Il s'agit d'une moyenne (f-mesure) des scores de rappel d'un système par rapport à l'autre. Les scores ne sont ni vraiment supérieurs aux valeurs de chaque système, ni vraiment inférieurs. Il semble donc que les systèmes sont aussi distants entre eux que de l'annotation de référence. Ils ne produisent donc pas les mêmes erreurs (l'accord n'augmente pas). De même, leurs sorties ne sont pas fondamentalement complémentaires (l'accord reste du niveau du score de rappel le plus bas).

Système LIA Les résultats de la condition *avecHC* nous permettent de distinguer les erreurs dues uniquement à la spécification du sens et à la résolution des références en contexte. L'étiquetage des liens référentiels (DER) est correcte à 86.5%. En ventilant ce résultat par rapport au type de liens référentiels (les spécifieurs), on remarque que les résultats sont très disparates : le taux d'étiquettes correctes atteint 95.2% pour les étiquettes *lienRef* sans spécifieur qui représentent 57% des liens référentiels du corpus ; il n'est que de 25% pour le spécifieur *ambigu*, associé à seulement 7.2% des liens du corpus. Cette faible représentativité de certains phénomènes pose problème aux méthodes probabilistes qui ont besoin d'un grand nombre d'exemples pour apprendre les modèles. L'identification des référents obtient des scores de précision/rappel convenables étant donné la simplicité des heuristiques mises en œuvre. Une analyse plus fine nous montre également que le système fait peu d'erreurs sur les références directes, qui sont aussi les plus nombreuses. Beaucoup d'erreurs se concentrent sur les liens *ambigus* et les liens contenant le spécifieur *inclusion* sont mieux traités que l'*exclusion*. Enfin notons que pour la condition de test *sansHC* la dégradation des résultats est limitée (-3% de precision, -12% de rappel) sachant que le taux d'erreur de l'annotation hors-contexte est de l'ordre de 20%. Ce dernier point souligne la robustesse de l'approche.

Système LORIA Le dépouillement des résultats du LORIA s'appuie sur une classification des erreurs IREF. Le premier résultat significatif est que 57% de nos erreurs proviennent de la résolution de la référence, alors que les 43% restants sont issus d'erreurs extérieures au module (projection hors-contexte, construction de la forme sémantique, analyse syntaxique ou lexicale). Par exemple si l'analyse

syntaxique oublie un hôtel parmi trois hôtels, le module de référence l'ignorera complètement de telle sorte que l'expression ultérieure "ces trois hôtels" échoue à trouver trois hôtels et ne retournera rien.

Nous avons ensuite raffiné les 57% d'erreurs de référence en vingt catégories classées en deux groupes. Nous distinguons les phénomènes non pris en compte (35%) et les réelles erreurs, problèmes et bugs (65%). Le premier groupe recouvre des cas complexes comme par exemple le générique "la chambre" en anaphore associative qui est mal géré en cas d'antécédent pluriel pour cause d'inconsistence logique. Le second groupe d'erreurs correspond à un fonctionnement anormal du module, par exemple rechercher le référent avec des contraintes qui n'auraient pas lieu d'être (comme dans "je voudrais plus de détails sur les hôtels" où l'on recherche "des hôtels avec plus de détails") ou encore une mauvaise structure domaniale. Nous pouvons conclure de cette analyse que le modèle des domaines de référence est très fin mais peu robuste. A la différence du système statistique du LIA très tolérant face aux erreurs, une erreur en début de dialogue peut entraîner ici une cascade d'erreurs.

Les résultats doivent cependant être compris en considérant la mesure d'évaluation qui a été adoptée. En effet cette dernière définit l'identité de deux référents comme une identité de description, traduisant alors mal le fait que deux référents peuvent être différents tout en se ressemblant. Par exemple, le référent de *"une chambre double à l'hôtel ibis paris"* et celui de *"une chambre double à l'hôtel lafayette paris"* sont similaires à 80% bien qu'ils représentent deux référents distincts. Le système du LIA est insensible au fait qu'il s'agisse de deux référents puisqu'il s'appuie directement sur les traits sémantiques de bon type dans le contexte antérieur, alors que le système du LORIA qui construit une représentation structurée des référents y est au contraire très sensible. Cette mesure ne permet alors d'évaluer que les capacités descriptives des référents, nécessaires dans un système de dialogue mais pas suffisantes. En effet la mesure ne permet pas de comparer les capacités référentielles pour lesquelles il est indispensable d'identifier avec précision le référent (en l'occurrence ne pas réserver une chambre dans le mauvais hôtel). Afin de considérer ces capacités, il est envisageable d'améliorer le couplage IREF pour qu'il ne couple que des référents décrits par les mêmes traits s'il s'agit d'entités nommées ou que des référents décrits par les mêmes traits discriminants s'il s'agit d'entités non-nommées (voir note 6).

7 Conclusion et perspectives

Cette étude présente la comparaison de deux approches très différentes pour résoudre le problème difficile de la résolution des références dans un contexte de dialogue. Les résultats obtenus permettent de mettre en avant les qualités et défauts des deux approches : bonne modélisation de phénomènes complexes mais faible tolérance aux erreurs pour le système symbolique ; bonne intégration avec les modules de décodage de parole et d'étiquetage conceptuel mais mauvaise prise en compte des références complexes pour le système probabiliste. A l'examen détaillé des résultats, la tâche de résolution des références semble être un domaine d'expérimentation prometteur pour étudier l'intégration des approches numériques et symboliques : en générant des listes d'hypothèses valuées, un système probabiliste peut être utilisé en entrée d'un système symbolique chargé de vérifier la cohérence des hypothèses générées, en supprimer certaines, afin de fournir les analyses les plus probables au gestionnaire de dialogue.

Remerciements

Nous tenons à remercier chaleureusement les annotatrices, Christelle AYACHE et Anne KUHN, pour la qualité de leur travail et leur participation enthousiaste et constructive à la définition du manuel d'annotation. Nous tenons également à remercier les relecteurs pour la pertinence de leurs commentaires.

Références

BONNEAU-MAYNARD H., AYACHE C., BECHET F., DENIS A., KUHN 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 *LREC'06*, Genoa.

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the french media dialog corpus. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal.

CHINCHOR N. & HIRSCHMANN L. (1997). MUC-7 coreference task definition, version 3.0. In *Actes de MUC-7*.

CORBLIN F. (1987). *Indéfini, Défini et Démonstratif*. Genève : Droz.

DENIS A., PITEL G. & QUIGNARD M. (2006a). A deep-parsing approach to natural language understanding in dialogue system : Results from a corpus-based evaluation. In *LREC 2006*, Genoa, Italy.

DENIS A., PITEL G. & QUIGNARD M. (2006b). Resolution of reference grouping in practical dialogues. In *SIGDial 2006*, Sydney, Australia.

GROSZ B. & SIDNER C. (1986). Attention, intention and the structure of discourse. *Computational Linguistics*, **12**, 175–204.

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.

LANDRAGIN F. & ROMARY L. (2004). Dialogue history modelling for multimodal human-computer interaction. In *Eighth Workshop on the Semantics and Pragmatics of Dialogue (Catalog'04)*.

MAYNARD H. & DEVILLERS L. (2000). A framework for evaluating contextual understanding. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, Beijing, Chine.

POPESCU-BELIS A., RIGOUSTE L., SALMON-ALT S. & ROMARY L. (2004). Online evaluation of coreference resolution. In *Proceedings of LREC 2004*.

REBOUL A., BALKANSKI C., BRIFFAULT X., GAIFFE B., POPESCU-BELIS A., ROBBA I., ROMARY L. & G. G. S. (1997). *Le projet CERVICAL : Représentations mentales, référence aux objets et aux événements*. Rapport interne, Loria-CNRS/Limsi, France.

REBOUL A. & GAIFFE B. (1999). Représentations mentales et référence.

REBOUL A. & MOESCHLER J. (1994). *Dictionnaire encyclopédique de la pragmatique*. Editions du Seuil.

SALMON-ALT S. (2001). *Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel*. PhD thesis, Université Henri Poincaré, Nancy.

SALMON-ALT S. & ROMARY L. (2004). Towards a reference annotation framework. In *Proceedings of LREC 2004*.

SERVAN C. & BECHET F. (2006). Décodage conceptuel et apprentissage automatique : application au corpus de dialogue homme-machine media. In *TALN*, Leuven.

VAN DEEMTER K. & KIBBLE R. (2000). On coreferring : Coreference in MUC and related annotation schemes. *Computational Linguistics*, **26**(4), 629–637.

Annotation précise du français en sémantique de rôles par projection cross-linguistique*

Sebastian PADÓ[1], Guillaume PITEL[2]
[1] Computerlinguistik – Université de la Sarre
[2] Équipe TALARIS, LORIA – INRIA
pado@coli.uni-sb.de, Guillaume.Pitel@loria.fr

Résumé. Dans le paradigme FrameNet, cet article aborde le problème de l'annotation précise et automatique de rôles sémantiques dans une langue sans lexique FrameNet existant. Nous évaluons la méthode proposée par Padó et Lapata (2005, 2006), fondée sur la projection de rôles et appliquée initialement à la paire anglais-allemand. Nous testons sa généralisabilité du point de vue (a) des langues, en l'appliquant à la paire (anglais-français) et (b) de la qualité de la source, en utilisant une annotation automatique du côté anglais. Les expériences montrent des résultats à la hauteur de ceux obtenus pour l'allemand, nous permettant de conclure que cette approche présente un grand potentiel pour réduire la quantité de travail nécessaire à la création de telles ressources dans de nombreuses langues.

Abstract. This paper considers the task of the automatic induction of role-semantic annotations for new languages with high precision. To this end we test the generalisability of the language-independent, projection-based annotation framework introduced by Padó and Lapata (2005, 2006) by (a) applying it to a new, more distant, language pair (English-French), and (b), using automatic, and thus noisy, input annotation. We show that even under these conditions, high-quality role annotations for French can be obtained that rival existing results for German. We conclude that the framework has considerable potential in reducing the manual effort involved in creating role-semantic resources for a wider range of languages.

Mots-clés : multilingue, FrameNet, annotation sémantique automatique, sémantique lexicale, projection d'annotation de rôles, rôles sémantiques.

Keywords: multilingual, FrameNet, automatic semantic annotation, lexical semantics, annotation projection, semantic roles.

1 Introduction

L'analyse sémantique de surface (en anglais, *shallow semantic parsing*) consiste à reconnaître les rôles sémantiques attribuables aux différents constituants d'un énoncé, sans décrire avec précision la sémantique interne de ces constituants. Les rôles sémantiques correspondent aux arguments des prédicats évoqués par certains mots, notamment les verbes. Ce type d'analyse présente un intérêt tout particulier dans les applications utilisant les informations sémantiques à grande échelle, telles que l'extraction d'information (Bouillon et *al.*, 2000 ; Surdeanu et *al.*, 2003), la traduction automatique (Boas, 2002) et les systèmes de question/réponse (Narayanan et Harabagiu, 2004).

* Ces travaux ont été financés par le fonds France-Berkeley (projet FR.FrameNet sous la responsabilité de C. J. Fillmore et L. Romary) et le DFG (Padó; bourse PI-154/9-2).

Le projet FrameNet (Fillmore et *al.*, 2003) peut jouer un rôle central dans les entreprises de ce type, en mettant à disposition une ressource lexicale de grande couverture, précisément articulée autour de la notion de rôle sémantique. Dans la sémantique des frames de Fillmore (1982) sur laquelle s'appuie FrameNet, le sens prédicatif est représenté à partir de *frames*[1] qui peuvent être considérées comme des représentations schématiques de situations. Dans ce cadre, les rôles sémantiques sont appelés *frame elements* (FEs), ils sont attachés de manière unique à une frame (on compte en moyenne 7 à 8 FEs par frame). Le projet FrameNet en lui-même consiste à construire une base de données mettant en relation les frames, les lemmes qui les évoquent (*Unités Lexicales ou ULs*) et les informations détaillées sur la réalisation de surface des Fes, sous la forme de cadres syntaxiques et grammaticaux et d'exemples d'annotations sur le British National Corpus. Ces informations ont créé un fort intérêt du côté de l'analyse sémantique de surface et ont rendu possible le développement d'analyseurs automatiques pour le texte libre (initiés par Gildea et Jurafsky, 2002) qui peuvent être utilisés pour les applications citées ci-dessus.

Frame: ARRIVING (Un THEME se rapproche d'un GOAL)			
Frame Elements	THEME	**The officer** approached the house **Amy** arrived home from school After **she** arrived home, …	**L'officier** s'approcha de la maison De l'école, **Amy** rentra à la maison Après **son** arrivée à la maison, …
	GOAL	The officer approached **the house** Amy arrived **home** from school He had arrived **there** from London	L'officier approcha_de **la maison** De l'école, Amy rentra **à la maison** Clarke est arrivé **là** de Londres
Unités Lexicales		approach.n, approach.v, arrival.n, arrive.v, come.v, crest.v, descend_(on).v, enter.v, entrance.n, entry.n, get.v, make it.v, make.v, reach.v, return.n, return.v, visit.n, visit.v	aborder.v, aboutir.v, accéder_à.v, approche.n, approcher.v, approcher_de.v, arriver.v, arrivée.n, atteindre.v, descendre_à.v, descendre_sur.v, entrer.v, entrée.n, gagner.v, parvenir.v, passer.v, rapprochement.n, regagner.v, rejoindre.v, rentrer.v, rentrée.n, retour.n, retourner.v, revenir.v, venir.v

Tableau 1: Exemple simplifié de *frame* dans le paradigme FrameNet (français et anglais)

Le tableau 1 illustre, pour la frame ARRIVING, certaines des informations contenues dans la base FrameNet (les ULs pour le français sont tirées d'une méthode semi-automatique de construction de lexique sémantique (Pitel, 2006)). Cette frame, qui modélise une situation où un objet en mouvement se rapproche d'un endroit, possède deux FEs principaux : le thème (THEME) et le but (GOAL). D'autres FEs secondaires lui sont rattachés, parmi lesquels : MODE_OF_TRANSPORTATION, CIRCUMSTANCES et GOAL_CONDITION[2]. Comme le montrent les exemples d'annotation, les mêmes FEs peuvent être évoqués par des constituants de différentes natures syntaxiques et grammaticales et ce sont ces informations qui seront particulièrement exploitées par les systèmes d'annotation sémantique automatique. Dans sa version 1.1 utilisée dans l'expérimentation que nous présentons, FrameNet décrit 513 frames liées à 7125 unités lexicales, avec en moyenne 7.5 FEs par frame.

Malheureusement, l'anglais est actuellement la seule langue dans laquelle une telle ressource existe à grande échelle. Un petit nombre de projets existent dans d'autres langues (allemand, espagnol et japonais principalement), mais ceux-ci n'ont pas encore atteint la maturité

[1] Afin d'éviter la confusion avec les autres usages de « cadre », nous conserverons les expressions usuelles anglaises lorsqu'elles comprennent le terme *frame*, et traduisons dans les autres cas.

[2] De nombreux exemples de frames ainsi que des documents sur FrameNet sont accessibles sur `http://framenet.icsi.berkeley.edu`

nécessaire à une exploitation automatique. La principale cause de ce déficit en ressources est le haut niveau d'exigence en temps et en attention requis pour l'annotation sémantique manuelle. Le fait que l'analyse sémantique de surface soit limitée à un petit nombre de langues est un obstacle important à son usage comme stratégie d'analyse générale pour le TAL. Il est donc impératif de concevoir des méthodes pour réduire l'effort nécessaire à l'amorçage de telles ressources. Dans cet article nous nous intéressons à l'induction des informations sémantiques sur les FEs, problème pour lequel (Padó et Lapata, 2005 ; 2006) – ci-après, P&L – ont suggéré l'utilisation de la *projection d'annotations*. Ce paradigme s'articule autour de l'exploitation de ressources parallèles entre des langues L1 et L2, où L1 est dotée en ressources de type FrameNet et L2 ne l'est pas. En supposant que l'on puisse obtenir une analyse sémantique pour le côté L1, la projection d'annotations consiste à *recopier* simplement les annotations de L1 sur le côté L2. et permet ainsi de réutiliser le travail manuel effectué sur L1.

P&L ont montré l'efficacité de cette méthode en utilisant FrameNet pour induire l'annotation de FEs en allemand, mais leur étude a été limitée sur deux aspects:

1. Seule une langue cible est considérée : l'allemand. Johansson et Nugues (2006) rapportent, avec une stratégie identique, un succès similaire pour le suédois. Cependant, le suédois et l'allemand étant deux langues germaniques, typologiquement proches de l'anglais, l'approche dans le cas général n'est pas validée.
2. P&L n'ont pris en compte que le cas où le côté L1 est annoté manuellement, ce qui ne permet pas de généraliser les résultats pour un passage à grande échelle. Bien que Johansson et Nugues (2006) l'aient fait pour une source automatique, les différences avec la méthode expérimentale de P&L ne permettent pas d'évaluer l'impact réel d'une annotation automatique.

Dans cet article, nous montrons que le paradigme de projection d'annotations des FEs se généralise au-delà du cas étudié dans P&L. Nous reproduisons les expériences réalisées dans P&L sur une langue romane, le français. Nous étendons de plus le champ d'investigation en réalisant une comparaison avec une projection à partir de FEs annotés automatiquement. Nous montrons que même cette situation permet d'obtenir des résultats de haute qualité pour le français. La structure de cet article est la suivante : nous présentons tout d'abord quelques travaux ayant exploré la projection cross-linguistique et décrivons la méthode de projection utilisée. Après avoir vérifier l'hypothèse fondamentale de parallélisme cross-linguistique des frames et Fes posée par cette méthode, nous présentons les résultats obtenus pour la projection de FEs de l'anglais vers le français. Nous concluons en évoquant les différentes pistes ouvertes par la projection cross-linguistique de *frame elements*.

2 Approches existantes

Le paradigme de la projection d'annotation a été introduit par Yarowsky et al. (2001), en utilisant un corpus parallèle pour adapter des outils monolingues (POS taggers, chunkers et analyseurs morphologiques) à des nouvelles langues. Le transfert effectif entre les langues a été rendu possible en utilisant les alignements de mots individuels entre les phrases, alignements qui peuvent aujourd'hui être obtenu automatiquement grâce à des outils comme GIZA++ (Och et Ney, 2003). Cette approche a été ensuite adaptée à d'autres niveaux de description linguistique, principalement pour la grammaire et la syntaxe. Par exemple, Hwa et al. (2002) ont projeté les informations de dépendance syntaxique de l'anglais au chinois.

La première tentative de transfert cross-linguistique d'informations sémantique a été présentée par Fung et Chen (2004) dans un projet de construction de FrameNet pour le chinois. Ceux-ci

exploitent les informations du FrameNet anglais en les mettant en relation avec des concepts tirés d'une ontologie en chinois, HowNet, sans d'ailleurs exploiter de corpus alignés. Leur stratégie requiert donc l'existence d'une grande ontologie pour la langue cible, ce qui n'est pas toujours disponible. L'approche de P&L se donne la même tâche, mais sans nécessiter une telle ressource, en se focalisant sur la projection d'annotations, moins gourmande en connaissances.

Cependant, la réussite de la projection d'annotations dépend en grande partie du *parallélisme cross-linguistique* de ces annotations. Comme la projection d'annotations basique consiste à copier l'information de l'annotation source, il y a erreur si l'annotation « idéale » de la cible n'est pas identique à l'annotation source. Le degré de parallélisme cross-linguistique est connu pour être très dépendant du niveau de description en question. Alors que Yarowsky et al. (2001) rapportent un parallélisme allant jusqu'à 85 % pour les étiquettes de partie du discours (anglais-français), Hwa et al. (2002) n'ont pas mesuré plus de 40 % de liens de dépendance syntaxique pouvant être projeté directement de l'anglais au chinois. P&L ont trouvé une bonne correspondance entre les frames et les FEs entre l'anglais et l'allemand, (voir section 4), ce qui donne une idée de la pertinence de l'approche. La question de la généralisation à d'autres langues, donc de l'évaluation du parallélisme sémantique, est cependant posée, car la proximité anglais-allemand est particulièrement importante.

3 Méthode de projection

Cette section décrit rapidement la méthode de projection proposée par P&L, qui définit un modèle général voulu indépendant de la langue, pour projeter les FEs de la langue source vers la langue cible. Contrairement aux études précédentes, P&L ont trouvé que les alignements mot à mot donnent de moins bons résultats pour la projection de FEs. Ces études concernaient la projection d'informations résidant au niveau des mots (Hwa et al., 2002) ou de syntagmes courts (Yarowsky et al., 2001), alors que les FEs peuvent s'étendre sur des syntagmes longs. Du fait des erreurs et des omissions dans les alignements automatiques de mots, la projection de longs constituants est délicate. Nous considérons donc comme P&L que les modèles de projections par alignement mot à mot (M) doivent être pris comme modèle de référence uniquement en l'absence d'informations plus riches.

D'autre part, P&L ont montré que la projection peut bénéficier énormément d'informations sur les constituants. Les transferts par constituants permettent à la fois d'espérer projeter les FEs sur des étendues pertinentes et d'avoir une meilleure robustesse que M, puisqu'un certain nombre d'erreurs d'alignement peut être compensé par les bons alignements. Par ailleurs, le fait que l'alignement soit recalculé au niveau des constituants permet des stratégies alternatives à l'alignement un à un. P&L proposent ainsi d'évaluer trois classes d'alignement de constituants : les alignements Total (T), Couvrant (C) et Exact (E) qui diffèrent dans la force des contraintes qu'ils imposent sur l'alignement.

Ces classes sont représentées figure 1 : T impose que chaque constituant source soit projeté au moins une fois, C impose en plus que chaque constituant cible soit lié à au moins une source et E impose que les constituants soient projetés un à un, introduisant éventuellement des constituants vides (ε). Un compromis doit être trouvé entre les alignements plus stricts comme E, qui peut corriger plus d'erreurs et un alignement plus souple comme T qui peut mieux modéliser les glissements dus à la traduction. Ce compromis pouvant être dépendant de la langue, nous comparons toutes ces différentes classes pour le français.

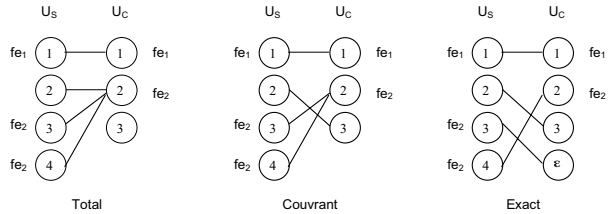


Figure 1: Modèles d'alignement de constituants (U_s et U_c sont les ensembles de constituants sources et cibles, fe_1 et fe_2 sont deux FEs)

Parallèlement aux différents modèles d'alignements, P&L proposent différents filtres pré- ou post-projection afin de réduire l'impact des erreurs d'alignements sur le résultat.[3]

1. Le filtre correctif de *convexité de couverture*, appliqué après projection, consiste à étendre artificiellement l'étendue d'un FE sur tous les éléments non annotés situés entre le premier et le dernier élément qui lui est attribuable. Johansson et Nugues (2006) étendent une telle heuristique en y ajoutant des caractéristiques spécifiques au suédois, ce qui leur permet d'obtenir d'excellents scores de projection.
2. Le filtrage de mots permet de corriger les erreurs dues à l'alignement : les mots grammaticaux souvent mal alignés et les nombreux mots laissés non alignés. La présence de ces erreurs dans un constituant peut le pénaliser inutilement dans le transfert. Pour palier ce problème, P&L proposent un filtre sur les mots grammaticaux (*MG*) et un autre sur les mots non alignés (*NA*).
3. Le *filtrage des non-arguments* (Arg) permet, à partir d'informations syntaxiques profondes, de ne pas prendre en compte les constituants qui ont peu de chances d'être des arguments du prédicat. L'efficacité de cette stratégie repose fortement sur la qualité de l'analyse syntaxique, afin que les constituants légitimes ne soient pas filtrés.

4 Parallélisme sémantique cross-linguistique

Nous avons argumenté dans la section 2 sur le fait que le parallélisme sémantique entre deux langues représente une borne maximum à la performance d'un système de projection de l'annotation. Pour cette raison, le parallélisme sémantique doit être évalué auparavant afin de déterminer la possibilité d'appliquer la projection à une langue particulière. Afin d'estimer ce parallélisme entre l'anglais et le français, nous avons produit un corpus annoté en sémantique des frames. Pour faire la comparaison avec les résultats de P&L, nous avons produit un sous-corpus en annotant les phrases correspondants au sous-corpus extrait pour leurs travaux. Ceci a été possible du fait que ce sous-corpus est tiré de Europarl (Koehn, 2005), corpus des minutes du parlement européen dans 11 langues. Le sous-corpus original a été extrait afin d'évaluer la projection entre l'allemand et l'anglais et est soumis à des contraintes spéciales qui introduisaient des biais que nous discuterons plus loin.

Etant donné la nature de l'expérimentation, consistant à mettre en relation une unique frame entre deux traductions, nous avons construit un guide pour l'annotation décrivant pour chaque phrase à annoter en français, le mot qui était le plus probablement le prédicat de la frame

[3] Nous ne testons pas les combinaisons possibles de filtres, suivant en cela P&L.

(trouvé par alignement à partir de la version anglaise annotée), ainsi que les frames potentiellement évoquées. En utilisant ce guide, deux annotateurs[4] ont produit une annotation du sous-corpus français de 1076 phrases, en utilisant 60 phrases pour la mise au point. Le sous-corpus était initialement analysé syntaxiquement par Syntex (Bourigault et *al.*, 2005) et a été annoté en utilisant l'outil SALTO (Burchardt et *al.*, 2006). L'évaluation de l'accord inter-annotateurs (avant adjudication) est résumée dans la colonne gauche du tableau 2 ; la colonne de droite présente les résultats de P&L. En général l'accord pour le français est élevé et correspond aux résultats pour l'allemand. Une des deux raisons auxquelles nous attribuons le score plus faible pour l'étendue des FEs est la nature plus fragmentaire de l'arbre syntaxique français (seuls 82 % des FEs français ont pu être assignés à des constituants uniques).

	Français	Allemand (P&L)
Acc. frames	0.87	0.87
Acc. FEs	0.89	0.95
Acc. étendue	0.72	0.83

Tableau 2: Accords inter-annotateurs pour les sous-corpus français (sur 500 phrases), comparés aux résultats de P&L pour l'allemand.

Le tableau 3 montre l'accord pour le français par rapport à l'anglais. Les résultats de P&L pour l'anglais-allemand sont donnés à titre de comparaison. La première ligne donne l'accord sur la frame choisie, la seconde sur les FEs. Nous avons trouvé que la correspondance cross-linguistique est quasiment identique pour les deux paires de langues. C'est le premier résultat important de notre étude, qui montre que le paradigme de projection est applicable aussi pour la paire anglais-français, malgré une plus grande distance typologique.

	Français/anglais	Allemand/anglais (P&L)
Corresp. frames	0.69	0.71
Corresp. FEs	0.88	0.91

Tableau 3: Correspondance interlingue des sous-corpus annotés

Il est intéressant de constater des différences dans la distribution des frames entre l'annotation française d'une part et les annotations anglaise et allemande d'autre part. Le tableau 4 donne le nombre de frames ayant un nombre d'annotation compris dans la fourchette décrite dans la colonne de gauche et entre parenthèses le total des annotations pour ces frames. Il montre des différences de répartition des frames dans l'annotation, qui s'expliquent par le fait que le sous-corpus a été initialement sélectionné pour maximiser les chances d'avoir des phrases avec des correspondances de frames entre l'allemand et l'anglais.

Nb annotations par frame	**Français**	**Allemand**	**Anglais**
25-160	8 *(418)*	13 *(578)*	9 *(447)*
10-24	20 *(315)*	14 *(228)*	25 *(389)*
1-9	**93 *(233)***	**45 *(133)***	**49 *(151)***
Total	121 *(966)*	73 *(987)*	83 *(987)*

Tableau 4: Distribution des frames en fonction du nombre d'annotations dans les trois sous-corpus: français, allemand et anglais.

[4] Nous remercions à ce propos Christiane Jadelot de l'ATILF pour son implication.

La distribution des frames est gonflée vers le bas, ce qui est une seconde raison pour le plus faible accord sur les étendues, les frames plus rares étant a priori plus difficiles à annoter.

5 Evaluation expérimentale

5.1 Conditions de l'expérimentation

Dans nos expériences, nous appliquons la méthode de projection au sous-corpus anglais-français que nous avons décrit dans la section 3. L'information sur les FEs est projetée de l'anglais sur le français. L'annotation manuelle en français sert de référence pour évaluer les annotations projetées. Nous comparons deux sources de projection différentes :

- *Condition 1 : annotation manuelle*. Cette annotation correspond à la configuration de P&L, dont le matériel est disponible. Comme nous l'avons évoqué, la pertinence de cette disposition est incertaine pour une application pratique, puisqu'en général, aucune annotation manuelle n'est disponible pour des corpus parallèles.
- *Condition 2 : annotation automatique.* Dans cette configuration nous avons utilisé un analyseur sémantique de surface de dernière génération (Giuglea et Moschitti 2006) entraîné sur les données de FrameNet 1.1, pour annoter les FEs sur le côté anglais du corpus[5]. Giuglea et Moschitti rapportent une précision de 85.2 % sur un jeu de données tiré de FrameNet ; dans notre évaluation avec l'annotation anglaise de référence de P&L, nous obtenons une f-mesure[6] de 65.1 (préc.: 78.1 %, rap.: 55.8 %). Les sources de différence sont les suivantes : (a) notre jeu de données « standard » n'inclut pas les traits de PropBank utilisés par Giuglea et Moschitti, (b) l'application à un corpus d'un autre domaine et (c) la couverture restreinte aux verbes, qui ne concernent que 87 % de notre corpus d'évaluation. Avec une évaluation limitée aux verbes, le système obtient un rappel de 62.4 %.

Afin de pouvoir rendre nos résultats comparables avec P&L, nous suivons leur démarche: le corpus parallèle est divisé en un corpus de développement et un corpus de test (50 % chacun). Dans les deux conditions, nous utilisons l'ensemble de développement pour comparer le modèle M de référence avec les modèles sur constituants, chacun combinant une classe d'alignement avec une procédure de filtrage. Les résultats des meilleurs modèles pour chaque alignement sont ensuite vérifiés sur l'ensemble de test. Toutes les évaluations utilisent l'alignement automatique intersectif produit par GIZA++.

5.2 Résultats

Modèle \ Filtre	∅	NA	MG	Arg
Mots	50.7 *(53.3/48.3)*	50.7 *(53.3/48.3)*	30.4 *(32.5/28.6)*	-
Total	53.5 *(57.2/50.2)*	57.8 *(68.1/50.2)*	45.6 *(60.5/36.6)*	64.1 *(71.4/58.1)*
Couvrant	**55.9** *(60.0/52.3)*	62.6 *(65.9/59.7)*	61.9 *(66.8/57.6)*	**64.2** *(71.5/58.3)*
Exact	54.7 *(60.9/49.6)*	**63.4** *(68.9/58.7)*	**62.3** *(69.4/56.6)*	60.6 *(84.2/47.3)*

Tableau 5: Evaluations dans l'ensemble de développement, source manuelle (condition 1)

[5] Nous remercions Ana-Maria Giuglea et Alessandro Moschitti pour l'accès à leur logiciel.
[6] F-mesure = (2×Rappel×Précision)÷(Rappel+Précision)

Nous présentons les résultats sur l'ensemble de développement dans les tableaux 5 (condition 1) et 6 (condition 2). Les résultats sur l'ensemble de test sont dans le tableau 7. Le format des mesures reproduites est : « F-mesure *(%Précision/%Rappel)* ». Les meilleures configurations globales d'après l'ensemble de développement sont en grisé, les meilleures f-mesures par filtre sont en gras, les meilleures f-mesures par modèle sont soulignées.

Modèle \ Filtre	∅	NA	MG	Arg
Mots	45.4 *(54.6/38.9)*	*45.4 (54.6/38.6)*	*28.1 (34.0/24.0)*	-
Total	47.9 *(58.4/40.6)*	51.3 *(69.7/40.6)*	42.8 *(66.3/31.6)*	**59.5** *(74.4/49.6)*
Couvrant	**51.7** *(62.9/43.9)*	57.6 *(68.8/49.6)*	56.7 *(69.5/47.9)*	**59.5** *(74.4/49.6)*
Exact	51.5 *(65.0/42.6)*	**58.3** *(71.7/49.1)*	**57.1** *(72.0/47.3)*	55.9 *(84.6/41.7)*

Tableau 6: Evaluations dans l'ensemble de développement, source automatique (condition 2)

Modèle	Condition 1		Condition 2	
Mots	∅ :	49.3 *(50.6/48.1)*	∅ :	45.4 *(54.6/38.9)*
Total	Arg :	62.7 *(68.3/57.9)*	Arg :	56.1 *(72.0/47.9)*
Couvrant	Arg :	63.0 *(68.8/58.3)*	Arg :	55.9 *(71.6/45.9)*
Exact	NA :	**63.1** *(66.2/60.3)*	NA :	**57.2** *(70.2/48.3)*

Tableau 7: Evaluation des projections dans l'ensemble de test, comparaison des modèles dans chaque condition (avec les meilleurs filtres selon l'ensemble de développement).

5.3 Comparaisons et discussion

Nous considérons tout d'abord les résultats sur l'ensemble de développement pour la condition 1, tableau 5. On constate que les modèles à base de constituants dépassent systématiquement la référence du modèle M, ce qui signifie que la segmentation est utile aussi pour le français. Les résultats de la condition 1 sont largement similaires aux résultats de P&L dans plusieurs aspects. La qualité globale est proche : le meilleur modèle sur le français (f=64.2 %) est seulement 3 % en dessous du score pour l'allemand obtenu par P&L (f=67.3 %). Les résultats sur le français sont encore plus favorables quand ils sont comparés à leur borne supérieure, qui est l'accord inter-annotateur (à défaut d'avoir une évaluation manuelle des projections) : ils sont à 6 % sous le plafond de 72 %, là où les résultats allemands sont 16 % sous l'accord à 83 %. Ensuite, on observe le même impact des procédures de filtrage. Les résultats montrent clairement un impact positif des filtres NA et Arg. En revanche, le filtre NA qui limite le bruit dû à l'alignement montre de meilleurs résultats quand le modèle est plus restrictif (Couvrant et Exact par rapport à Total), alors que le filtre Arg améliore nettement la précision mais fait baisser le rappel, et favorise les modèles relativement moins restrictifs (Total et Couvrant par rapport à Exact). Ces observations montrent que les modèles d'alignement et les filtres de P&L sont pertinents au-delà de la paire de langues pour laquelle ils ont été créés. La généricité de ce paradigme doit être prise en compte pour comparer nos résultats avec ceux de Johansson et Nugues (2006), qui annoncent une f-mesure de 82.0 (préc. : 84.0, rap. : 81.0) pour leur projection. Ils utilisent des heuristiques spécifiques au suédois[7], qui devront être

[7] L'accord inter-annotateur de leur corpus de référence n'étant pas connu,et comme celui-ci ne comporte que 150 phrases, il n'est pas possible de faire une comparaison des deux méthodes.

recréées pour chaque nouvelle langue cible et sont probablement plus difficile à identifier pour des langues plus distantes que l'anglais et le suédois, qui sont des langues très proches (Koehn 2005).

Ensuite, nous nous intéressons aux résultats utilisant une source automatique (condition 2), qui sont exposés dans le tableau 6 pour l'ensemble de développement. On note que le passage de la version manuelle à celle annotée automatiquement conduit à une perte de performance allant de seulement 3 à 7 points sur la f-mesure. Cette différence correspond assez bien à la différence observée entre les annotations manuelles (f=72 %, d'après l'accord inter-annotateurs), et l'analyseur sémantique de surface (f=65 %). Une comparaison des tableaux 5 et 6 montre que les caractères de l'annotation automatique (haute précision, rappel bas) sont très clairement reproduits dans les propriétés de la projection : alors que la source automatique amène une nette chute du rappel, la précision reste constante voire s'améliore par rapport à la source manuelle. Cet apparent paradoxe vient du fait que la projection à partir de la source automatique ne tente pas de projeter certain FEs difficiles, par exemple ceux qui s'étendent sur plusieurs constituants, simplement parce que les FEs n'ont pas été annotés par le système automatique sur l'anglais. En somme, les résultats de la condition 2 montrent la possibilité d'utiliser des analyseurs sémantiques de surface comme entrée de la projection, ce qui est indispensable pour appliquer ce paradigme à grande échelle. En plus, les annotations obtenues sont de grande précision, caractéristique essentielle pour la création de ressources.

Les résultats sur l'ensemble de test sont exposés dans le tableau 7. Moins riches d'enseignement que les résultats de l'ensemble de développement et par manque de place, nous n'en ferons qu'une synthèse courte. Une baisse de 2 à 3 % est observée en moyenne sur la f-mesure, par rapport à l'ensemble de développement, mais cette différence peut s'expliquer par une variation aléatoire naturelle sur le partitionnement du corpus (P&L obtiennent des résultats légèrement meilleurs avec le même découpage). Les différences entre les méthodes basées sur constituants ne sont pas significatives, mais le meilleur modèle (E+NA) pour le français est le même que le meilleur modèle pour l'allemand trouvé par P&L.

6 Conclusion et perspectives

Dans cette étude nous avons montré la possibilité de produire des annotations sémantiques de grande précision pour le français en appliquant la démarche proposée par Padó et Lapata (2005, 2006). Cette démonstration procède en trois étapes : (1) la vérification que le parallélisme français-anglais du point de vue de la sémantique des frames est suffisant pour permettre la projection de l'annotation, (2) la démonstration que les résultats obtenus par P&L pour la paire anglais-allemand se reproduisent bien pour la paire anglais-français aussi bien en terme de performance absolue que du point de vue des effets des différents filtrages – ce qui montre la généricité dans la démarche suivie et valide une fois encore l'aspect multilingue de FrameNet – et (3) la démonstration que même l'utilisation d'une source bruitée provenant d'une annotation automatique résulte en une annotation qui, en particulier, montre une haute précision. D'ores et déjà, une analyse plus approfondie de nos données a montré que 124 phrases du corpus ont une annotation parfaitement identique à l'annotation manuelle de référence. Ces résultats ont de fortes chances de s'améliorer avec des avancées dans les technologies d'alignement, d'analyse syntaxique et sémantique de surface.

Avec un tel résultat de référence, relativement facile à obtenir (alignement automatique, analyseur syntaxique et corpus aligné avec l'anglais), on peut très raisonnablement envisager d'entraîner dès maintenant des annotateurs automatiques de FEs en français à moindre coût. Les résultats obtenus par Johansson et Nugues (2006) nous indiquent que des techniques spécifiques à la langue cible, qui éliminent les FEs non plausibles projetés durant l'induction

de l'analyseur sémantique, permettent de produire un annotateur automatique ayant une performance à peine dégradée par rapport aux versions natives. Ces bons résultats autorisent à envisager une méthode de construction de ressources fondée sur un travail itératif, utilisant à la fois un outil automatique pour produire des annotations et une phase de révision manuelle pour obtenir un corpus de bonne qualité lexicographique. En effet, l'effort requis pour démarrer une telle ressource est rendu abordable par cette approche, même si une évaluation plus précise du coût de correction par rapport au coût de production reste à faire afin de valider totalement cette démarche.

Références

BOAS H.C. (2002). Bilingual FrameNet dictionaries for machine translation. Actes de *LREC 2002*, pp. 1364–1371, Las Palmas, Iles Canaries.

BOUILLON P., FABRE C., SÉBILLOT P., JACQMIN L. (2000). Apprentissage de ressources lexicales pour l'extension de requêtes, *Traitement Automatique des Langues*, 41:2, pp. 367—393.

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P., OZDOWSKA S. (2005). Syntex, analyseur syntaxique de corpus, Actes de *TALN 2005*, Dourdan, France.

BURCHARDT A., ERK K., FRANK A., KOWALSKI A., PADO S. (2006). SALTO – A Versatile Multi-Level Annotation Tool. Actes de *LREC 2006*, Gênes, Italie.

FILLMORE C.J. (1982). Frame Semantics. *Linguistics in the Morning Calm*, pp. 111—38. Seoul.

FILLMORE C.J., JOHNSON C.R., PETRUCK M.R. (2003). Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

FUNG P., CHEN B. (2004). BiFrameNet: Bilingual frame semantics resources construction by cross-lingual induction. Actes de *COLING 2004*, pp. 931—935, Genève, Suisse.

GILDEA D., JURAFSKY D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

GIUGLEA A.-M., MOSCHITTI A. (2006). Semantic role labeling via FrameNet, VerbNet and PropBank. Actes de *COLING/ACL 2006*, pp. 929—936, Syndey, Australie.

HWA R., RESNIK P., WEINBERG A., KOLAK O. (2002). Evaluation translational correspondance using annotation projection. Actes de *ACL 2002*, pp. 392—399, Philadelphia.

JOHANSSON R., NUGUES P. (2006). A FrameNet-based Semantic Role Labeler for Swedish. Actes de *COLING/ACL 2006* Main Conf. Poster Sessions, pp. 436—443, Sydney, Australie.

KOEHN P. (2005). Europarl: A parallel corpus for statistical machine translation. Actes de *MT Summit X*. Phuket, Thaïlande.

NARAYANAN S., HARABAGIU S. (2004). Question answering based on semantic structures. Actes de *COLING 2004*, pp. 693–701, Genève, Suisse.

OCH F. J., NEY H. (2003). A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1):19-51.

PADÓ S., LAPATA M. (2005). Cross-lingual projection of role-semantic information. Actes de *HLT/EMNLP* 2005, Vancouver, Canada.

PADÓ S., LAPATA M. (2006). Optimal Constituent Alignment with Edge Covers for Semantic Projection. Actes de *COLING/ACL 2006*, pp. 1161—1168, Sydney, Australie.

PITEL G. (2006). Using bilingual LSA for FN annotation of French text from generic resources. Workshop on Multilingual Annotation : Theory and Applications, Saarbrücken.

SURDEANU M., HARABAGIU S., WILLIAMS J., AARSETH P. (2003). Using predicate-argument structures for information extraction. Actes de *ACL 2003*, pp. 8—15, Sapporo, Japon.

YAROWSKY D., NGAI G., WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. Actes de *HLT 2001*, pp. 161—168. San Francisco.

Session
Acquisition

Élaboration automatique d'un dictionnaire de cooccurrences grand public

Simon CHAREST, Éric BRUNELLE, Jean FONTAINE, Bertrand PELLETIER
Druide informatique inc.
1435, rue St-Alexandre, bureau 1040
Montréal (Québec) H3A 2G4, Canada
developpement@druide.com

Résumé. Antidote RX, un logiciel d'aide à la rédaction grand public, comporte un nouveau dictionnaire de 800 000 cooccurrences, élaboré essentiellement automatiquement. Nous l'avons créé par l'analyse syntaxique détaillée d'un vaste corpus et par la sélection automatique des cooccurrences les plus pertinentes à l'aide d'un test statistique, le rapport de vraisemblance. Chaque cooccurrence est illustrée par des exemples de phrases également tirés du corpus automatiquement. Les cooccurrences et les exemples extraits ont été révisés par des linguistes. Nous examinons les choix d'interface que nous avons faits pour présenter ces données complexes à un public non spécialisé. Enfin, nous montrons comment nous avons intégré les cooccurrences au correcteur d'Antidote pour améliorer ses performances.

Abstract. Antidote is a complete set of software reference tools for writing French that includes an advanced grammar checker. Antidote RX boasts a new dictionary of 800,000 co-occurrences created mostly automatically. The approach we chose is based on the syntactic parsing of a large corpus and the automatic selection of the most relevant co-occurrences using a statistical test, the log-likelihood ratio. Example sentences illustrating each co-occurrence in context are also automatically selected. The extracted co-occurrences and examples were revised by linguists. We examine the various choices that were made to present this complex data to a non-specialized public. We then show how we use the co-occurrence data to improve the performance of Antidote's grammar checker.

Mots-clés : antidote, cooccurrences, collocations, corpus, analyseur, correcteur.
Keywords: antidote, co-occurrences, collocations, corpus, parser, grammar checker.

1 Introduction

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. Parmi les dix dictionnaires de l'édition RX figure un nouveau dictionnaire de cooccurrences, constitué essentiellement automatiquement au moyen d'outils de traitement de la langue à couverture large. Le dictionnaire de cooccurrences d'Antidote a la particularité d'être destiné au grand public, pour qui il établira souvent le premier contact avec la notion de cooccurrence.

Par *cooccurrence*, nous entendons la présence simultanée et statistiquement significative, dans un corpus, de deux unités linguistiques en relation syntaxique. Notre définition correspond aux *cooccurrences relationnelles* d'Evert (2005), aussi appelées *cooccurrences syntaxiques*, par opposition aux *cooccurrences positionnelles*, dont les mots apparaissent simplement fréquemment dans une certaine proximité.

Notre concept de cooccurrence englobe des combinaisons lexicales dont le degré de figement est variable : nous y incluons à la fois des combinaisons libres (*entendre un cri*), des combinaisons semi-figées ou *collocations* au sens strict (*pousser un cri*) et des *locutions* figées courantes (*cri du cœur*) ou terminologiques (*cri primal*).

Comme Antidote comporte déjà un dictionnaire de locutions, nous avions d'abord pensé nous restreindre aux *collocations*, mais deux motifs nous ont incités à élargir notre cible :

1. La frontière entre collocation et locution figée, d'une part, et entre collocation et combinaison libre, d'autre part, n'est pas toujours nette. Par exemple, dans certains cas limites, il n'est pas clair si une combinaison de mots acquiert vraiment un sens différent de la composition des sens de chacune de ses parties.

2. Pour obtenir la description la plus complète possible de l'usage d'un mot, il nous est apparu plus intéressant de présenter toutes les combinaisons les plus fortes de ce mot, incluant les locutions figées et les combinaisons libres statistiquement significatives.

Tutin (2005) examine l'intérêt d'un dictionnaire de cooccurrences et constate les lacunes des dictionnaires généralistes en la matière. Comme outil de consultation, un dictionnaire de cooccurrences énumère les contextes d'usage d'un mot. En production de texte, il complète le dictionnaire de synonymes pour retrouver le mot juste ou le tour idiomatique. Il aide enfin à l'apprentissage du français langue seconde, car les locutions et collocations, qui reflètent des emplois figés ou semi-figés, sont difficilement prédictibles pour un locuteur étranger.

Le dictionnaire de cooccurrences d'Antidote RX a été extrait automatiquement à partir d'un corpus de 500 millions de mots. L'analyseur syntaxique de haut niveau 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. Le présent article décrit le processus d'élaboration de ce dictionnaire.

2 Travaux antérieurs

Les dictionnaires de cooccurrences en français sont peu nombreux. Un pionnier fut Ulysse Lacroix avec *Les mots et les idées : dictionnaire des termes cadrant avec les idées*, dont la première édition remonte à 1931. Plus récemment, le *Dictionnaire des cooccurrences* de Beauchesne (2001) recense environ 150 000 cooccurrences. Le site Web *Dictionnaire des collocations* de Rodriguez, lancé en 2004, affichait 31 400 combinaisons à la fin de 2006. Ces trois dictionnaires ont été élaborés à partir d'une collecte manuelle des cooccurrences. Le *Dictionnaire combinatoire du français* de Zinglé et Brobeck-Zinglé (2003) a été créé à partir d'un corpus dont ont été tirées 65 000 expressions (34 000 en version imprimée).

Parmi les dictionnaires universitaires, mentionnons le *Dictionnaire explicatif et combinatoire du français contemporain* (Mel'čuk et coll., 1984, 1988, 1992, 1999), dont une version électronique simplifiée développée par Mel'čuk et Polguère peut être consultée sur le site *DiCouèbe*. On peut aussi consulter, sur le site de l'Équipe de recherche en syntaxe et sémantique (ERSS), la base lexicale distributionnelle *Les voisins de Le Monde*, construite automatiquement à partir d'un corpus comprenant l'ensemble des articles du quotidien *Le Monde* sur une période de dix ans (1991-2000), soit environ 200 millions de mots.

L'extraction automatique de cooccurrences à partir d'un corpus a fait l'objet de nombreux articles. Plusieurs traitent de cooccurrences positionnelles, basées sur la seule proximité des mots à l'intérieur d'un intervalle donné (par ex. de 5 positions). Parmi les travaux qui mentionnent un traitement syntaxique, soulignons ceux de Lin (1998), qui utilise un parseur et filtre les cooccurrences en fonction de l'information mutuelle ; ceux de Kilgarriff et Tugwell (2001) pour *Word Sketch*, un module du *Sketch Engine* accessible en ligne, où les cooccurrences sont extraites du *British National Corpus* à l'aide d'un étiqueteur et d'une grammaire de type *pattern matching*, puis filtrées selon une mesure intégrant l'information mutuelle ; et enfin les travaux du Laboratoire d'Analyse et de Technologie du Langage (LATL) de l'Université de Genève (Seretan, Wehrli, 2006), qui utilisent le parseur multilingue *Fips* et filtrent les résultats en employant la mesure du rapport de vraisemblance.

3 Méthodologie

3.1 Constitution d'un corpus

Le matériau brut de notre dictionnaire étant un corpus, il est essentiel que celui-ci soit de grande taille, afin de refléter un large éventail d'usage des mots, des plus fréquents aux plus rares. Il faut aussi varier les styles d'écriture ainsi que les domaines des textes (voir Tableau 1). Nous récoltons en outre des écrits de diverses régions de la francophonie dans le but d'extraire des cooccurrences propres aux locuteurs de chacune de ces régions.

Domaine	Proportion	Exemples de sources
Littéraire	30 %	Gallica, Projet Gutenberg, Les Éditions Québec Amérique
Journalistique	40 %	Le Devoir, Voir, L'Express, Libération, La Tribune de Genève
Autres	30 %	Wikipédia, CyberSciences, LégiFrance, Université de Montréal

Tableau 1: principales sources du corpus

Le Web, mine quasi inépuisable de textes de toute sorte, est la principale source de notre corpus. Mais le Web a aussi ses défauts. Par exemple, il est fréquent de trouver des phrases qui apparaissent de manière récurrente dans plusieurs pages d'un même site. On trouve aussi des phrases, provenant de citations ou de dépêches journalistiques, qui se retrouvent sur plusieurs sites différents, parfois telles quelles, parfois légèrement reformulées. Ces phrases récurrentes faussent les statistiques de fréquence. Il a donc fallu identifier et éliminer automatiquement les phrases identiques ou trop similaires.

Au final, nous avons constitué un corpus de 500 millions de mots, répartis sur 25 millions de phrases. C'est de ce matériau brut que nous avons extrait nos cooccurrences.

3.2 Extraction des cooccurrences

L'analyseur d'Antidote, fruit de plus de 10 années de développement intensif, reconnait un large éventail de structures syntaxiques du français. Nous en avons créé une version adaptée, optimisée pour analyser en traitement distribué la masse énorme du corpus et en extraire les cooccurrences. Sur une grappe de 15 ordinateurs utilisant la technologie de distribution XGrid d'Apple, il lui a fallu 1100 heures pour analyser les 500 millions de mots du corpus.

L'analyseur effectue une analyse en dépendance et génère des arbres syntaxiques complets, desquels les cooccurrences sont extraites directement. Lorsque plusieurs analyses sont trouvées pour une même phrase, l'arbre le plus probable, selon la pondération de l'analyseur, est choisi. Nous avons sélectionné les relations syntaxiques les plus pertinentes pour un dictionnaire de cooccurrences (voir Tableau 2). Pour cette première version, nous n'avons considéré que les cooccurrences à deux membres.

Categ 1	Categ 2	Relation	Proportion	Exemple
Nom	Adjectif	Épithète	18 %	*jeune fille*
Nom	Nom	Apposition	1 %	*site internet*
Nom	Nom/Verbe	Complément du nom	25 %	*coup d'œil*
Verbe	Adverbe	Modificateur	4 %	*aller loin*
Verbe	Nom	Sujet	12 %	*le vent souffle*
Verbe	Nom	Complément direct	10 %	*jouer un rôle*
Verbe	Nom	Autres compléments[1]	26 %	*perdre de vue*
Verbe	Nom/Adj	Attribut	1 %	*retenir prisonnier*
Verbe	Verbe	Complément verbal	1 %	*entendre parler*
Adjectif	Adverbe	Modificateur	1 %	*gravement malade*
Adjectif	Nom	Complément de l'adjectif	1 %	*âgé de x ans*

Tableau 2 : principales relations extraites

Au-delà des relations syntaxiques directes, le système s'efforce d'extraire des cooccurrences en franchissant des relations plus profondes. Par exemple :

- Coordinations : « De temps en temps cette clameur et ce bruit redoublaient. » → *la clameur redouble* ; *le bruit redouble*
- Relatives : « Or, sa réouverture est cruciale pour acheminer l'aide humanitaire dont a désespérément besoin la population. » → *avoir besoin d'aide*
- Agents : « Nous laissons à nos lecteurs le soin de deviner quel genre de surprise cette chute apporterait aux habitants. » → *le lecteur devine*

[1] Les « autres compléments » incluent les compléments indirects (COI) et les compléments adverbiaux ou circonstanciels.

- Collectifs : « Il a ouvert une infinité de routes, toutes raboteuses, qu'il a fallu ensuite aplanir. » → *ouvrir la route*

Nous avons considéré certaines locutions verbales figées (p. ex. *faire face*, *laisser place*, *avoir besoin*) comme des verbes à part entière, ce qui nous a permis d'extraire des cooccurrences comme *faire face aux défis*, *laisser place au doute* et *avoir besoin d'aide*.

Outre les deux mots formant la cooccurrence, nous notons certaines données morpho-syntaxiques qui définissent la distribution de ses emplois. Nous retenons ainsi, pour chaque cooccurrence, les types des déterminants, le genre et le nombre de chaque mot, et la position relative de ceux-ci. Ces données déterminent la formulation la plus fréquente de la cooccurrence, qui sera utilisée notamment pour l'affichage.

Au total, plus de 17 millions de cooccurrences distinctes ont été extraites, chacune apparaissant en moyenne 4 fois dans le corpus.

3.3 Sélection des cooccurrences

3.3.1 Sélection automatique

Un test statistique permet de faire un premier tri des cooccurrences. Le test le plus simple serait d'utiliser directement les fréquences, en ne retenant que les cooccurrences d'une fréquence minimale donnée. Mais un tel test peut difficilement s'appliquer à l'ensemble des mots d'une langue, car les fréquences varient énormément d'un mot à l'autre. Plusieurs autres *mesures d'association* ont toutefois été proposées pour quantifier la force d'une combinaison de mots, dont le rapport de vraisemblance (*log-likelihood ratio*), l'information mutuelle, le test *t* et le test du khi-carré. Pour une description approfondie de diverses mesures d'association, voir (Evert, 2005), Manning & Schütze (1999) ou Dunning (1993).

Selon Dunning (1993), l'information mutuelle et le test *t* ont tendance à surestimer la force des combinaisons de faible fréquence ou dont un des composants est rare. En revanche, le rapport de vraisemblance s'appuie sur des fondements statistiques solides pour comparer directement l'importance d'évènements rares et d'évènements fréquents. De plus, selon Orliac (2004), il s'agit de la mesure la plus apte à isoler les collocations d'un ensemble de combinaisons. Pour ces raisons, nous avons choisi le rapport de vraisemblance comme mesure de la force de nos cooccurrences.

Nous l'avons mentionné à la section 3.1, la diversité du corpus a un impact direct sur les cooccurrences extraites. Lors de nos premiers essais, certaines cooccurrences à priori peu intéressantes mais présentant une valeur de force anormalement élevée provenaient souvent d'un seul et même texte ou site Web. Nous avons dû élaborer une euristique pour tenir compte de la dispersion d'une combinaison à travers plusieurs sources. Plutôt que la fréquence brute, nous employons la somme des racines carrées des fréquences pour chaque source. Ainsi, une combinaison apparaissant 9 fois dans une même source aura le même poids que si elle apparaissait une fois dans 3 sources distinctes.

Après avoir calculé la force des combinaisons candidates, nous avons filtré les moins intéressantes en fixant empiriquement un seuil pour chaque type de relation syntaxique. Cet ensemble de seuils a ainsi retranché 93,5 % des 17 millions de cooccurrences extraites. De

plus, les cooccurrences de fréquence 1 (apax) et celles formées de mots banals, comme le verbe « être » et certains adverbes (« très », « trop »…), ont été filtrées automatiquement, retirant un autre 1,5 %. Au total, la sélection automatique retient donc un peu plus de 850 000 cooccurrences « brutes ».

3.3.2 Révision manuelle

Les cooccurrences brutes ont ensuite fait l'objet d'une révision « manuelle » par des linguistes, afin de vérifier la qualité des résultats et de rejeter les cooccurrences jugées indésirables. Ces rejets concernent :

- Des cooccurrences mal analysées : des phrases complexes peuvent donner du fil à retordre à l'analyseur et lui faire générer des analyses incorrectes. Par exemple, *président du trésor*, mauvaise analyse de *président du Conseil du trésor*.

- Des cooccurrences incomplètes : l'extracteur ne considérant que deux mots principaux par cooccurrence, il lui arrive de générer une cooccurrence incomplète comme *emploi à temps*, à laquelle il manque un adjectif essentiel (*plein* ou *partiel*). Parfois, le linguiste peut y remédier en insérant manuellement certains éléments récurrents (*nager en délire > nager en plein délire*) ou paires binaires de tels éléments (*blague de gout > blague de bon/mauvais gout*).

- Des cooccurrences peu intéressantes : certaines cooccurrences offrent peu d'intérêt pour un dictionnaire malgré leur relative fréquence. Par exemple, des combinaisons libres avec un gentilé (*ambassadeur américain*, *français*, etc.), ou un verbe de sens très générique (*avoir un chien*, *un livre*, etc.).

- Des cooccurrences délicates : cooccurrences de caractère offensant, expressions de registre vulgaire, anglicismes condamnés par Antidote, etc.

Un autre objet de la révision humaine est de vérifier la formulation choisie de manière automatique et de rectifier, le cas échéant, certains attributs morphosyntaxiques. Par exemple, le linguiste peut remplacer un article par un possessif (*rencontrer sur le passage > rencontrer sur son passage*), insérer une négation (*la pluie discontinue > la pluie ne discontinue pas*), modifier le genre (*abandonné par son mari > abandonnée par son mari*), etc.

Après huit mois de révision, 5 % seulement des cooccurrences ont ainsi été rejetées manuellement. Nous obtenons donc 800 000 cooccurrences qui constituent le dictionnaire final.

3.4 Choix des exemples

Nous avons choisi d'illustrer chaque cooccurrence par des exemples réels tirés du corpus. Un algorithme vorace sélectionne un ensemble minimal de phrases qui serviront d'exemples. L'algorithme s'efforce de minimiser la longueur totale des phrases tout en maximisant leur « qualité ». La qualité d'une phrase tient compte de plusieurs paramètres, dont :

- sa longueur (ni trop courte ni trop longue) ;

- la qualité de la source d'où elle provient ;

- le nombre de fautes identifiées par l'analyseur ;
- le nombre de noms propres (l'idée étant de minimiser les noms propres).

L'algorithme tente aussi de maximiser la diversité des sources, pour éviter de choisir, pour une même cooccurrence, plusieurs phrases du même ouvrage, du même auteur, ou du même site Web. En revanche, l'algorithme essaie de réutiliser le plus possible une même phrase pour illustrer plusieurs cooccurrences distinctes, afin de minimiser la taille totale des données. Enfin, les phrases trop similaires pour une même cooccurrence sont identifiées et coupées.

Parmi 11 millions de phrases candidates, 870 000 phrases sont ainsi sélectionnées, représentant plus de 2 millions d'exemples.

De ce nombre, 300 000 phrases ont été identifiées automatiquement pour être révisées par des linguistes, dans le but de rejeter les phrases jugées indésirables selon certains critères, parmi lesquels :

- Présence d'erreurs non détectées par l'analyseur ; phrases de mauvaise qualité, trop « orales », parsemées d'anglais, en ancien français, etc.
- Présence de mots offensants, inconvenants, vulgaires ; propos non neutres sur des sujets délicats.
- Mention d'une personne non publique ; présence de coordonnées, de numéros de téléphone, etc. ; phrases à caractère publicitaire.

Environ le quart des phrases révisées ont été ainsi rejetées, puis remplacées et révisées à nouveau, en un processus itératif de sélection automatique et de révision manuelle.

4 Présentation des cooccurrences

Présenter efficacement une masse de plusieurs centaines de cooccurrences à un utilisateur non spécialisé présente plusieurs défis. La Figure 1, ouverte sur le début des cooccurrences du mot *grève*, donne un aperçu des choix que nous avons faits et que nous discutons ci-dessous.

Nous estimons que des cooccurrences complètes ont plus d'impact et sont plus faciles à lire que celles où le mot-vedette est omis ou remplacé par un symbole. De plus, nous jugeons qu'il est utile de présenter les cooccurrences avec leurs attributs morphosyntaxiques réels. Nous affichons donc *déclencher une grève*, *mettre fin à la grève*, *voter en faveur d'une grève*, ce qui correspond aux formulations les plus fréquentes. D'autres formulations ont pu être rencontrées dans le corpus (*déclencher la grève*, *déclencher cette grève*, *les grèves ont été déclenchées*…) ; les exemples d'occurrences peuvent donner un aperçu de cette variabilité. Lorsqu'une cooccurrence est sélectionnée, des exemples de phrases tirées du corpus s'affichent en effet dans le panneau de droite, avec surlignement des mots de la cooccurrence.

Dans le cas d'un mot polysémique, comme *grève*, les cooccurrences sont regroupées sous chacun des sens, affichés en vert, afin de faciliter la consultation[2]. Un triangle de dévoilement permet de réduire les listes sous chaque sens afin d'accéder rapidement au sens désiré.

Les cooccurrences sont ensuite classées par relation syntaxique. Un histogramme discret illustre la force relative de chaque cooccurrence. Au départ, seules les cinq cooccurrences les plus fortes de chaque relation sont affichées, pour donner une meilleure vue d'ensemble. Le lien *et x autres,* à la fin de chaque liste, permet de dévoiler la suite. Un triangle de dévoilement permet de réduire chacune des listes individuellement.

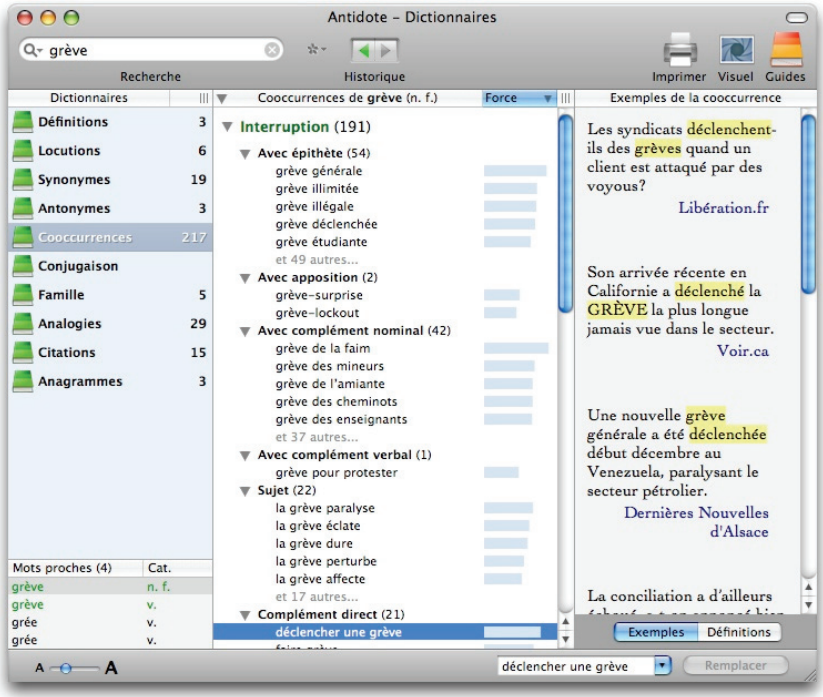


Figure 1: interface du dictionnaire de cooccurrences d'Antidote RX

[2] Pour l'affichage, les cooccurrences dont les bases sont polysémiques ont été manuellement classées sous les divers sens de la base (environ 4000 sens sous 900 bases). Par exemple, les cooccurrences de grève au sens d'« interruption de travail » ont été séparées de celles de grève au sens de « plage ».

5 Utilisation dans le correcteur

Le dictionnaire de cooccurrences d'Antidote RX n'est pas seulement un outil de consultation, il est aussi utilisé par le correcteur pour raffiner l'analyse syntaxique, détecter et corriger des fautes sémantiques et éliminer certaines alertes inutiles.

Les cooccurrences guident l'analyseur en lui permettant d'éliminer des branchements syntaxiques moins probables si un branchement plus fort est détecté. Ce faisant, l'analyseur purement symbolique devient un analyseur hybride intégrant aussi des notions statistiques. Cette avancée a permis de réduire d'environ 10 % le nombre d'arbres syntaxiques produits, augmentant du même coup la vitesse d'analyse.

Les erreurs de nature sémantique, comme **tache ingrate* ou **perpétrer la tradition*, ne peuvent être repérées par la seule analyse syntaxique. Mais comme ces erreurs proviennent souvent de la confusion entre homonymes ou paronymes, il devient possible, en consultant les cooccurrences, de déterminer si un homonyme ou un paronyme est statistiquement plus probable et de proposer la correction le cas échéant. Antidote peut corriger 25 000 confusions de ce type.

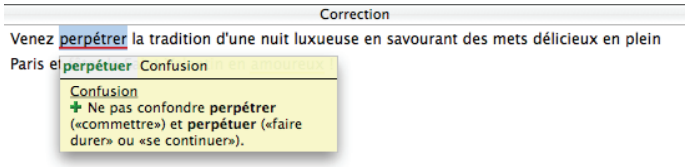


Figure 2 : correction d'une erreur de nature sémantique

Inversement, si le correcteur trouve deux mots en relation de cooccurrence forte, il peut taire certaines alertes de nature sémantique qu'il aurait autrement générées. Par exemple, sur un mot qui peut être un faux ami (ou anglicisme sémantique), comme *digital* au sens de *numérique*, le correcteur n'alertera pas l'utilisateur si le mot est employé dans un contexte de cooccurrence forte, comme *empreinte digitale*.

6 Conclusion

Nous avons vu comment a été élaboré un dictionnaire de cooccurrences commercial, au moyen d'outils de traitement automatique de la langue (TAL) à couverture large appliqués à un vaste corpus du français. Nous avons vu également comment l'analyseur syntaxique utilisé pour extraire les cooccurrences en a lui-même tiré profit pour améliorer ses performances d'analyse et de correction.

Le dictionnaire de cooccurrences d'Antidote RX est aujourd'hui entre les mains de milliers d'utilisateurs. Plusieurs ont manifesté leur satisfaction, et même noté que, de tous les outils d'Antidote RX, le dictionnaire de cooccurrences est déjà devenu celui qu'ils consultent le plus souvent. D'autres ont été agréablement surpris par une correction « sémantique » pertinente. Ces résultats montrent que le grand public peut profiter pleinement des avancées du TAL, et qu'il est prêt à accueillir de nouveaux outils linguistiques avancés et inédits.

Remerciements

Nous tenons à remercier Mala Bergevin, Jean Saint-Germain, Jasmin Lapalme, Marie-Hélène Gaudreault, Sophie Campbell, Sara-Anne Leblanc, Ophélie Tremblay, Guy Lapalme, Alain Polguère et toute l'équipe des druides sans qui Antidote RX n'aurait pu voir le jour.

Références

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

EVERT S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

KILGARRIFF A., TUGWELL D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. *Proc. Collocations workshop, ACL 2001, Toulouse*, 32-38.

LIN, D. (1998). Extracting Collocations from Text Corpora. *First Workshop on Computational Terminology, COLING-ACL '98, Montréal*, 57-63.

MANNING C., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge : The MIT Press.

ORLIAC B. (2004). *Automatisation du repérage et de l'encodage des collocations en langue de spécialité.* Thèse de doctorat présentée à l'Université de Montréal.

SERETAN V., WEHRLI E. (2006). Accurate collocation extraction using a multilingual parser. *Proceedings of COLING-ACL 2006, Sydney, Australia*, 953-960.

TUTIN A. (2005). Le dictionnaire de collocations est-il indispensable ? *Revue française de linguistique appliquée,* X-2, 31-48.

Dictionnaires

BEAUCHESNE, J. (2001). *Dictionnaire des cooccurrences.* Montréal : Guérin.

LACROIX, U. (1931). *Les mots et les idées. Dictionnaire des termes cadrant avec les idées.* Paris/Bruxelles.

MEL'ČUK I. *et coll.* (1984, 1988, 1992, 2000). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I, II, III, IV*. Montréal : Les Presses de l'Université de Montréal.

MEL'ČUK I., POLGUÈRE A. (Sous presse). *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Louvain-la-Neuve : Duculot.

ZINGLÉ H., BROBECK-ZINGLÉ M.-L. (2003). *Dictionnaire combinatoire du français. Expression, locutions et constructions.* Paris : La Maison du Dictionnaire.

DiCouèbe, dictionnaire en ligne de combinatoire du français : http://olst.ling.umontreal.ca/dicouebe

Dictionnaire des collocations : http://www.tonitraduction.net

*Les voisins de Le Mond*e : http://w3.univ-tlse2.fr/erss/voisinsdelemonde

Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux

Didier SCHWAB[1], Lim LIAN TZE[1], Mathieu LAFOURCADE[2]
[1] Computer-Aided Translation Unit (UTMK)
School of Computer Sciences, Universiti Sains Malaysia
Penang, Malaysia
[2] TAL-LIRMM, Université Montpellier II – CNRS
161 rue ada, 34392 Montpellier Cedex 5, France
{didier,liantze}@cs.usm.my, lafourcade@lirmm.fr

Résumé. Fréquemment utilisés dans le Traitement Automatique des Langues Naturelles, les réseaux lexicaux font aujourd'hui l'objet de nombreuses recherches. La plupart d'entre eux, et en particulier le plus célèbre *WordNet*, souffrent du manque d'informations syntagmatiques mais aussi d'informations thématiques (« *problème du tennis* »). Cet article présente les vecteurs conceptuels qui permettent de représenter les idées contenues dans un segment textuel quelconque et permettent d'obtenir une vision continue des thématiques utilisées grâce aux distances calculables entre eux. Nous montrons leurs caractéristiques et en quoi ils sont complémentaires des réseaux lexico-sémantiques. Nous illustrons ce propos par l'enrichissement des données de *WordNet* par des vecteurs conceptuels construits par émergence.

Abstract. There is currently much research in natural language processing focusing on lexical networks. Most of them, in particular the most famous, *WordNet*, lack syntagmatic information and but also thematic information (« *Tennis Problem* »). This article describes conceptual vectors that allows the representation of ideas in any textual segment and offers a continuous vision of related thematics, based on the distances between these thematics. We show the characteristics of conceptual vectors and explain how they complement lexico-semantic networks. We illustrate this purpose by adding conceptual vectors to *WordNet* by emergence.

Mots-clés : *WordNet*, vecteurs conceptuels, informations lexicales, informations thématiques.

Keywords: *WordNet*, conceptual vectors, lexical information, thematic information.

1 Introduction

Originellement issus des travaux de Ross Quillian sur la psycholinguistique à la fin des années 60 (Quillian, 1968), les réseaux lexicaux sont toujours aujourd'hui au centre des recherches en Traitement Automatique des Langues Naturelles. Ils sont utilisés dans de nombreuses tâches (désambiguisation lexicale (Mihalcea *et al.*, 2004)) ou applications du domaine (traduction automatique avec les réseaux multilingues comme Papillon (Mangeot-Lerebours *et al.*, 2003) ou

(Knight & Luk, 1994), recherche d'informations ou classification de textes (Harabagiu & Chai, 1998)). La plupart de ces réseaux, et spécifiquement le plus célèbre d'entre eux *WordNet* (Fellbaum, 1988), souffrent du manque d'informations syntagmatiques mais aussi d'informations concernant le domaine d'usage des termes ou du moins les termes thématiquement associés. Il n'y a ainsi aucune relation directe entre des termes comme ‹*teacher*›-‹*student*› (‹*enseignant*›-‹*étudiant*›) et ‹*boat*›-‹*port*› (‹*bateau*›-‹*port*›). Ce phénomène a été nommé « *Problème du tennis* » [(Fellbaum, 1988), p. 10] lorsqu'il a été remarqué qu'il fallait chercher les équivalents de ‹*balle*›, ‹*raquette*› et ‹*court*› à différents endroits de la hiérarchie.

Depuis quelques années, l'équipe de traitement automatique des langues (TAL) du LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) travaille sur une formalisation de la projection de la notion linguistique de champ sémantique dans un espace vectoriel, les vecteurs conceptuels. Ils permettent de représenter les idées contenues dans un segment textuel quelconque et permettent d'obtenir une vision continue des thématiques utilisées grâce aux distances calculables entre eux.

Dans cet article, nous présentons les vecteurs conceptuels et en particulier leur version émergente. Nous montrons leurs caractéristiques et en quoi ils sont complémentaires des réseaux lexico-sémantiques. Nous illustrons ce propos par une expérience menée à Penang en Malaisie qui a consisté à enrichir les données de *WordNet* de vecteurs conceptuels par émergence.

2 Réseaux lexico-sémantique : l'exemple de *WordNet*

Principe et lacunes. *WordNet* est une base de données lexicale pour l'anglais développée sous la direction de George Armitage Miller par le *Cognitive Science Laboratory* de l'université de Princeton (États-Unis d'Amérique). Il se veut représentatif du fonctionnement de l'accès au lexique mental humain.

WordNet est organisé en ensembles de synonymes appelés synsets. À chaque synset correspond un concept. Le sens des termes est décrit dans *WordNet* par trois moyens : (1) leur *définition* ; (2) le *synset* auquel ce sens est rattaché ; (3) les *relations lexicales* qui unissent entre eux les synsets. On trouve parmi ces relations, l'hyperonymie, la méronymie et l'antonymie.

La version 2 de *WordNet* compte 152059 termes ce qui constitue une couverture relativement large de la langue anglaise. Dans les premières versions de *WordNet*, les relations lexicales ne connectent que les termes de même morphologie. Il y a donc une hiérarchie pour les noms, une pour les adjectifs, une pour les verbes et enfin une dernière pour les adverbes.

Dans (Harabagiu *et al.*, 1999), les auteurs de *WordNet* (alors à sa version 1.6) relèvent six faiblesses dans la construction de leur réseau : (1) le manque de liens entre les hiérarchies ; (2) le nombre limité de relations entre termes traitant du même sujet ; (3) le manque de relations morphologiques ; (4) l'absence de relations thématiques ; (5) l'absence de certains sens de mots ; (6) le manque d'uniformisation et de cohérence dans les définitions. Si les points 3, 5 et 6 ne nous intéressent pas dans cet article, nous allons montrer l'apport des vecteurs conceptuels pour la résolution des autres, tous trois formant le problème du tennis.

Expériences cherchant à résoudre le problème du tennis. Dans cet article, nous nous intéresserons uniquement à la version 2.1 de *WordNet* qui était la dernière disponible au moment

où nous avons réalisé nos expériences. Une nouvelle version (3.0) est sortie en Décembre 2006 mais elle ne semble pas comporter de réelles améliorations par rapport à la version précédente pour ce qui nous intéresse ici.

Depuis la version 2, des relations comme *derivationally related form* (formes dérivationnelles) permettent de lier des adjectifs à des verbes ou des adjectifs à des noms. De même, les synsets peuvent se voir attribuer un domaine d'usage. Toutefois, ces données semblent encore en nombre trop restreint pour être suffisamment pertinentes. Des relations typiques comme ‹*teacher*›-‹*student*› (‹*enseignant*›-‹*étudiant*›) ‹*boat*›-‹*port*› (‹*bateau*›-‹*port*›) ou ‹*doctor*›-‹*hospital*› (‹*docteur*›-‹*hôpital*›), pourtant souvent indispensables à une tâche de désambiguïsation lexicale, ne s'y trouvent toujours pas et le nombre restreint d'indications thématiques comme l'est le domaine ne permet pas de compenser ce défaut. Plusieurs solutions ont été proposées pour résoudre tout ou partie de ce problème.

Avec *eXtended WordNet*, (Harabagiu *et al.*, 1999) propose de désambiguïser l'ensemble des définitions de *WordNet* de façon semi-automatique. L'idée est, pour chaque définition, de dire quel est le sens utilisé pour chacun des termes. On peut ensuite comparer deux synsets et évaluer leur similarité. Nous verrons que nous utilisons ces informations pour fabriquer les vecteurs conceptuels de cette expérience. D'autres eux aussi rajoutent des informations aux synsets. Ainsi, (Agirre *et al.*, 2001) ajoutent des signatures lexicales issues de corpus taggés ou du Web. En revanche, d'autres cherchent plutôt à augmenter le nombre d'arcs existants. (Stevenson, 2002), par exemple, combine différentes métriques pour créer des arcs entre synsets à partir de leur définition et d'un thésaurus. (Ferret & Zock, 2006) utilisent eux un réseau de coocurrences pour extraire des relations typiques comme celles présentées dans un paragraphe précédent.

On le voit, toutes ses propositions ont en commun d'appartenir en particulier au domaine du discret. La nôtre est d'introduire une représentation continue des idées contenues dans le réseau, les vecteurs conceptuels.

3 Les vecteurs Conceptuels

Nous présentons ici les points fondamentaux à comprendre sur les vecteurs conceptuels. Nous revenons sur le mode de construction classique des vecteurs conceptuels, c'est-à-dire tels qu'ils ont été étudiés au LIRMM depuis 1997[1], à partir d'un ensemble de concepts choisis *a priori*. Nous expliquons dans cette partie certaines notions de base qui nous seront utiles pour présenter ensuite la construction par émergence, c'est à dire sans concepts prédéfinis.

Principe Généraux. Nous représentons les aspects thématiques des segments textuels (documents, paragraphes, syntagmes, etc.) par des vecteurs conceptuels, une formalisation de la projection de la notion linguistique de champ sémantique dans un espace vectoriel. À partir d'un ensemble de notions élémentaires dont nous faisons l'hypothèse, les concepts[2], il est possible de construire des vecteurs dont chaque composante correspond à un concept et est positive. Par exemple, le vecteur de l'item lexical ‹*vie*›, qui fusionne tous les sens de ‹*vie*›, peut être projeté sur les concepts suivants (les CONCEPT⌈*intensité*⌋ sont ordonnés par valeurs décroissantes de l'intensité) : V ‹*vie*› = (VIE⌈0.7⌋, NAISSANCE⌈0.48⌋, ENFANCE⌈0.46⌋, MORT⌈0.43⌋, VIEILLESSE⌈0.41⌋, ...).

[1] Voir les articles de l'équipe dans les précédentes éditions de cette conférence ou (Schwab, 2005).

[2] Dans notre expérimentation sur le français nous utilisons (Larousse, 1992) qui défini 873 concepts.

La construction des vecteurs conceptuels se fait à partir de définitions extraites de diverses sources (dictionnaires, listes de synonymes, indexations manuelles, ...). Cette méthode d'analyse construit, à partir de vecteurs conceptuels déjà existants et de nouvelles définitions, de nouveaux vecteurs.

Distance angulaire. La comparaison entre deux vecteurs se fait grâce à la distance angulaire D_A. Pour deux vecteurs conceptuels A et B, $D_A(A,B) = arccos(Sim(A,B))$ où Sim est $Sim(X,Y) = \cos(\widehat{X,Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$. Intuitivement, cette fonction constitue une évaluation de la *proximité thématique* et en pratique la mesure de l'angle entre les deux vecteurs. Empiriquement, nous estimons que pour une distance $D_A(X,Y) \leq \frac{\pi}{4}$ (45°), X et Y sont thématiquement proches et partagent plusieurs concepts. Pour $D_A(X,Y) \geq \frac{\pi}{4}$, la proximité thématique est considérée comme faible et aux alentours de $\frac{\pi}{2}$ (90°), X et Y n'ont aucune relation. Nous obtenons, par exemple, les angles suivants :

D_A(V(‹*fourmilier*›), V(‹*fourmilier*›))=0 (0°) D_A(V(‹*fourmilier*›), V(‹*mammifère*›))=0.36 (21°)
D_A(V(‹*fourmilier*›), V(‹*animal*›))=0.45 (26°) D_A(V(‹*fourmilier*›), V(‹*quadrupède*›))=0,42 (24°)
D_A(V(‹*fourmilier*›), V(‹*train*›))=1.18 (68°) D_A(V(‹*fourmilier*›), V(‹*fourmi*›))=0,26 (15°)

Le premier résultat a une interprétation directe, ‹*fourmilier*› ne peut être plus proche d'autre chose que de lui même. Le fait qu'un ‹*fourmilier*› soit un ‹*mammifère*› explique le deuxième résultat. Un ‹*fourmilier*› n'a que peu de rapport avec un ‹*train*› ce qui explique l'angle plus important. Dans le dernier exemple, l'angle peu important entre ‹*fourmilier*› et ‹*fourmi*› se comprend si on se rappelle que D_A est une distance thématique et non une distance ontologique. L'examen de la définition de fourmilier, « *mammifère qui se nourrit de fourmis* », explique le résultat.

Le voisinage thématique, une vision continue de la thématique. La fonction de voisinage thématique permet de connaître les items lexicaux voisins d'un item lexical donné. On définit $\mathcal{V}$ la fonction de voisinage qui renvoie les k items les plus proches en termes de distance angulaire D_A d'un texte Z dans une base vectorielle. Soit
$|\mathcal{V}(D_A, Z, k)| = k \qquad \forall X \in \mathcal{V}(D_A, Z, k), \quad \forall Y \notin \mathcal{V}(D_A, Z, k), \quad D_A(\mathrm{X}, \mathrm{Z}) \leq D_A(\mathrm{Y}, \mathrm{Z})$
Par exemple, les 7 termes proches et ordonnés par distance thématique croissante du nom ‹*mort*› peuvent être :

$\mathcal{V}(D_A$, ‹*mort*›, 7) = (‹*mort*› 0) (‹*meurtre*› 0.367) (‹*tueur*› 0.377) (‹*âge de la vie*› 0.481) (‹*tyrannicide*› 0.516) (‹*tuer*› 0.579) (‹*mort :adj*› 0.582)

La méthode de voisinage peut être utilisée lors de l'apprentissage des vecteurs conceptuels pour vérifier la cohérence globale de la base ou en phase d'exploitation pour trouver le meilleur mot à utiliser dans un énoncé. Ainsi, elle constitue un nouvel outil pour accéder aux mots et à leur sens, complémentaire à ceux décrits dans (Zock, 2002) comme la forme, la morphologie ou la navigation dans un grand réseau lexical. La fonction de voisinage permet ainsi une navigation dans le domaine du continu contrairement aux réseaux sémantiques qui ne permettent qu'une navigation discrète.

Somme vectorielle. Soient X et Y deux vecteurs, leur *somme vectorielle normée* V est définie par : $\vartheta^2 \rightarrow \vartheta \ : V = X \oplus Y \quad | \quad V_i = \frac{X_i + Y_i}{\|X+Y\|}$ où ϑ est l'ensemble des vecteurs conceptuels, V_i (resp X_i, Y_i) représente la i-ème composante du vecteur V (resp. X, Y).

La somme vectorielle normée de deux vecteurs donne un vecteur équidistant en termes d'angle des deux premiers vecteurs. Il s'agit en fait d'une moyenne des vecteurs sommés. En tant

qu'opération sur les vecteurs conceptuels, on peut donc voir la somme vectorielle normée comme l'union des idées contenues dans les termes.

Soient X et Y deux vecteurs, leur *produit terme à terme normalisé* V est défini par : $\vartheta^2 \rightarrow \vartheta \ : V = X \otimes Y \quad | \quad v_i = \sqrt{x_i y_i}$ L'opérateur $\otimes$ peut être interprété comme un opérateur d'intersection entre vecteurs. Si l'intersection entre deux vecteurs est le vecteur nul, alors ils n'ont rien en commun. Du point de vue des vecteurs conceptuels, cette opération permet donc de sélectionner les idées communes à un ensemble de termes.

Construction des vecteurs par émergence. L'approche par émergence s'affranchit de tout thésaurus et vecteurs de concept comme base de départ. Seule d la taille du vecteur est fixée *a priori*. Le mode de construction des vecteurs est identique au modèle classique à la différence que si un des vecteurs entrant dans la somme est inexistant, car non encore calculé, alors ce vecteur est tiré au hasard. Le processus de calcul est itéré jusqu'à convergence de chaque vecteur.

Comme nous le montrons de façon plus détaillée dans (Lafourcade, 2006), il y a un certain nombre d'avantages à utiliser ce modèle. Le premier d'entre eux est de pouvoir choisir librement la quantité de ressources que l'on souhaite utiliser en choisissant la taille des vecteurs de façon appropriée. Pour donner une idée de l'importance de ce choix, une base de 500000 vecteurs de dimension 1000 fait environ 2Go, de taille 2000, 4Go, ... Comme il ne serait pas alors ni raisonnable ni facile de définir une jeu de concept de la taille choisie, autant chercher une approche nous permettant de nous en passer. De plus, ce qui peut sembler un pis-aller ou au mieux un compromis, s'avère un avantage car la densité lexicale dans l'espace des mots calculés par émergence est bien plus constante que dans un espace où les concepts sont précalculés. En effet, les ressources (les dimensions de l'espace) ont tendance à être harmonieusement distribuées en fonction de la richesse lexicale.

4 Modélisation hybride du sens : vecteurs conceptuels et réseaux lexicaux

4.1 Apport des réseaux lexicaux aux vecteurs conceptuels

Les distances utilisées sur les vecteurs, comme le montre (Besançon, 2001), mettent en exergue les composantes communes et/ou les composantes distinctes. Si nous utilisons en particulier la distance angulaire, c'est que ses caractéristiques mathématiques, sa simplicité à comprendre et à interpréter linguistiquement ainsi que son efficacité en termes de temps de calcul en font un bon outil. Quelle que soit la distance choisie, utilisée sur ce type de vecteur (représentant des idées, des concepts plutôt que des termes cooccurrents), elle est d'autant plus faible que les vecteurs des objets lexicaux qui en sont les arguments sont dans un champ sémantique proche (en isotopie selon la terminologie de Rastier (Rastier, 1985)).

Dans le cadre d'une analyse sémantique comme celle qui nous intéresse ici, nous l'utilisons pour tirer profit des informations mutuelles contenues dans les vecteurs conceptuels pour faire de la désambiguïsation lexicale sur des mots qui ont des sens situés dans un champ sémantique proche. Ainsi, « *Zidane a marqué un but* » peut être désambiguïsée grâce aux idées communes concernant le sport tandis que « *L'avocat a plaidé à la cour* » peut l'être grâce à celles concer-

nant la justice. De même, en ce qui concerne les rattachements prépositionnels, les vecteurs peuvent permettre dans « *Il voit la fille avec un télescope.* » de rattacher « *avec un télescope* » au verbe ‹*voir*› grâce aux idées communes sur la vision.

En revanche, les vecteurs conceptuels ne peuvent pas aider à résoudre des cas où les termes mis en jeu sont dans des champs sémantiques différents. On remarquera même qu'une analyse ne reposant que sur eux peut conduire à de gros contre-sens. Par exemple, dans la phrase « *L'avocat a mangé un fruit* », ‹*avocat*› ne peut être interprété que comme le fruit et non comme l'auxiliaire de justice. Ces limites des vecteurs conceptuels ont été expérimentalement montrées pour l'analyse sémantique sur des algorithmes à fourmis dans (Lafourcade & Guinand, 2006).

Il aurait fallu que des connaissances comme « *un avocat est un être humain* » et « *un être humain mange* » puissent être identifiées, ce qui n'est donc pas possible avec des vecteurs conceptuels seuls. Les vecteurs conceptuels seuls ne sont ainsi pas suffisants pour exploiter certaines instances de fonctions lexicales dans les textes et un réseau lexical peut donc aider à pallier ces manques. Des publications antérieures ont montré la nécessité de cette approche hybride : (Schwab *et al.*, 2002) pour les antonymies, (Lafourcade & Prince, 2003) pour les génériques et les hyperonymes. (Schwab, 2005) étend cette constatation à toute relation susceptible d'aider à la résolution d'une analyse sémantique.

4.2 Apport des vecteurs conceptuels aux réseaux lexicaux

S'ils bénéficient d'une précision certaine, le rappel des réseaux est bien moins fort. Il est, en effet, difficile de penser que l'on pourrait représenter toutes les relations entre les termes. En effet, comment considérer deux termes qui sont dans le même champ sémantique ? Ils peuvent très bien ne pas se trouver dans le réseau car ils ne seraient pas forcément reliés par un des arcs "classiques". Envisager l'introduction d'arcs de type *champ sémantique*, poserait à nos yeux deux problèmes dus au caractère flou et flexible de cette relation :

– le premier est lié à l'idée de la relation que se fait le concepteur de la base, à quel moment considère-t'il que deux synsets sont dans le même champ sémantique ? Dans un cas défavorable, on aurait très peu de ces arcs tandis que dans un cas opposé, on pourrait se trouver avec une explosion combinatoire du nombre d'arc ;
– le second problème, plus fondamental, est lié à la représentation elle même. Comment envisager de représenter par un élément discret une relation floue donc du domaine du continu ?

Ainsi, le domaine du continu offert par les vecteurs conceptuels offre des flexibilités que le domaine du discret offert par les réseaux ne peut donner. Il permet de pouvoir rapprocher des mots sur des idées peu importantes mais pourtant communes à deux objets.

Avec cette approche hybride - vecteurs conceptuels, réseau lexical - nous proposons de combiner des informations de nature complémentaire. Les vecteurs conceptuels et l'opération de distance thématique par leur nature peuvent pallier le faible rappel intrinsèque aux réseaux lexicaux tandis que ces derniers peuvent permettre de désambiguïser les cas qui sont dans un champs sémantique différent contrairement aux vecteurs conceptuels. Les défauts des uns sont ainsi compensés par les qualités des autres ce qui fait des vecteurs conceptuels et des réseaux lexicaux des outils complémentaires.

5 Expérience sur *WordNet* : utilisation des données

5.1 Exploitation des définitions

Le projet *eXtended WordNet* (Mihalcea & Moldovan, 2001) est mené à la *Southern Methodist University* de Dallas au Texas et vise deux objectifs : (1) désambiguïser l'ensemble des termes utilisés dans les définitions des synsets, c'est-à-dire indiquer quels sont les synsets employés dans la définition ; (2) Transformer ces définitions en forme logique pour permettre plus facilement les calculs.

Ces données ont été réalisées de façon semi-automatique en utilisant les informations du réseau[3], des distances entre définitions ou bien les informations sur le domaine. Ces données sont en partie contrôlées à la main et le taux de précision de plus de 90%.

Pour les vecteurs conceptuels, nous avons utilisé ces données sous forme logique car elles permettent de repérer les éléments les plus importants de la définition, en particulier le genre. Le calcul se fait ainsi sur un arbre en dépendances fabriqué à partir de cette définition prétraitée pour enlever le métalangage difficilement exploitable pour une analyse thématique. Dans nos explications, nous allons prendre pour exemple la forme logique de la définition de fourmi.

ant :NN(x1) -> social :JJ(x1) insect :NN(x1) live :VB(e1, x1, x3) in :IN(e1, x2) organized :JJ(x2) colony :NN(x2)

Elle est organisée en 3 ensembles : x1 = {*social*, *insect*}, x2 = {*organised*, *colony*} et e1 = {*live*}. Ce dernier ainsi que *in* permettent de hiérarchiser les ensembles. Le vecteur de chacun des ensembles est calculé en faisant la somme vectorielle de l'élément le plus porteur de sens de cet ensemble (verbes, VB ; noms, NN) et de la moitié des adjoints (adverbes, RB ; adjectifs, JJ). Le calcul du vecteur global se fait ensuite par somme vectorielle pondérée des différents ensembles dans l'arbre en commençant par la partie la plus basse. Ce mode de calcul permet de considérer de façon prépondérante le genre sur les autres termes de la définition et de façon plus générale les têtes sur leurs dépendants syntaxiques. La figure 1 synthétise ce calcul. Aucun prédicat n'étant dans l'ensemble x3, il n'apparaît pas sur le schéma.

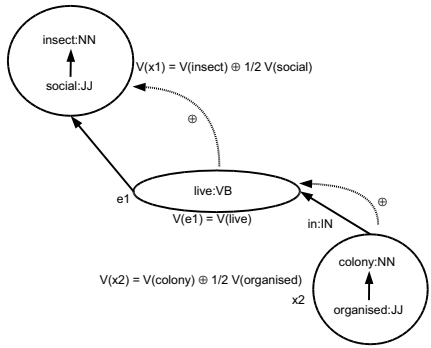


FIG. 1 – Construction du vecteur conceptuel de la définition de fourmi

[3]Par exemple, pour une définition aristotélicienne (en genre et différences), si le genre a un sens qui est aussi un hyperonyme du synset défini, on considère que ce sens est celui utilisé dans la définition.

5.2 Exploitation des relations

L'exploitation des relations se fait à deux niveaux : (1) pour la construction des vecteurs, elles permettent de fabriquer de manière complémentaire aux définitions le vecteur d'un synset ; (2) pour éviter les phénomènes de regroupement d'ensembles distincts.

5.2.1 Construction des vecteurs

La construction d'un vecteur conceptuel est effectuée pour chaque nœud du réseau par simple somme pondérée des vecteurs des nœuds reliés. Soit un nœud N relié a k nœuds $N_1 \dots N_k$, le vecteur de N, $V(N)$ sera égal à $p_1V(N_1) + p_2V(N_2) + \dots + p_kV(N_k)$ où p_i est le poids du i-ème nœud. Le vecteur somme est ensuite normalisé.

Cette approche entraîne naturellement une agglomération des vecteurs. Il est donc nécessaire d'augmenter le contraste d'un vecteur à la suite de son calcul. Pour ce faire, on calcule le coefficient de variation[4] de V. Si ce dernier ne se situe pas a 10% du CV moyen alors le vecteur subit une opération non linéaire d'amplification (la mise à une puissance n de chaque composante puis normalisation), et ce de façon itérée jusqu'à l'obtention d'un coefficient de variation dans la fourchette acceptable. Cette dernière a été estimée à partir des valeurs obtenues dans les expériences avec concepts prédéfinis.

5.2.2 Problème du regroupement d'ensembles distincts

Un dernier problème potentiel est que les vecteurs de deux ensembles distincts (à la fois au sens du réseau lexical et de la thématique) de termes peuvent occuper la même région de l'espace. L'approche du calcul se faisant par activation et les vecteurs étant tirés au hasard à l'initialisation rien n'empêche que cela se produise par accident. Il est donc nécessaire de "séparer" les vecteurs proches mais correspondant pourtant à des parties très différentes du réseau lexical et de la thématique.

La détection de ce phénomène se fait par scrutation du voisinage d'un vecteur conceptuel. Si parmi ses n premiers voisins, la densité de mots n'ayant rien à voir avec le mot étudié est importante alors une action de séparation doit être entreprise.

Cette action de séparation consiste à plonger l'ensemble du réseau dans un champs où les nœuds ont tendance à se repousser. En s'inspirant directement de la physique, une force de répulsion en $1/d^2$ est calculée itérativement entre les nœuds. Pour un nœud donné, on peut ainsi calculer un vecteur déplacement qui va l'éloigner des nœuds dont il se trouve trop près. Les nœuds ne se rapprochant pas par voisinage thématique (lors de la première phase du calcul) mais se trouvant proches "par accident" finissent ainsi naturellement par se séparer.

[4] Le coefficient de variation CV est donné par la formule $\frac{EC(V)}{\mu(V)}$ avec EC(V) l'écart type du vecteur V et $\mu(V)$ la moyenne arithmétique des composantes de V.

6 Conclusion

Dans cet article, nous avons présenté les vecteurs conceptuels construits par émergence. Nous avons montré en quoi ils peuvent aider à résoudre le « *problème du tennis* » de par leur caractère complémentaire aux réseaux lexico-sémantiques dont l'exemple le plus courant dans les recherches actuelles est *WordNet*. En effet, le rappel des réseaux est faible, ils ne permettent pas facilement de représenter le champs sémantique contrairement aux vecteurs tandis que ces derniers ne sont pas suffisants pour représenter des relations comme l'hyperonymie ou la méronymie.

Notre proposition est de tirer profit de cette complémentarité en ajoutant à *WordNet* des vecteurs conceptuels construits à partir des définitions et des relations contenues dans cette base. La méthode proposée ici tient du domaine du continu contrairement à l'ensemble des méthodes que nous avons étudiées dans la littérature qui, elles, font partie du domaine du discret (ajout d'arcs pour les relations, de symboles sur le domaine, etc.).

Nous avons conscience que cette méthode ne permet seulement que de résoudre une partie du « *problème du tennis* ». En effet, les vecteurs conceptuels ne permettent pas d'exhiber les rapports collocationnels non-thématiques entre items. Il s'agit essentiellement des relations qu'Igor Mel'čuk modélise avec ses fonctions lexicales syntagmatiques (Mel'čuk *et al.*, 1995) comme l'intensification (« *peur bleue* » ; *Magn* (‹*peur*›) = ‹*bleue*›), la dégradation (« *lait tourne* » ; *Degrad* (‹*lait*›) = ‹*tourner*›) ou bien encore le confirmateur (« *argument valable* » ; *Ver* (‹*argument*›) = ‹*valable*›). Comme le remarque (Ferret & Zock, 2006), ces relations font partie de celles qu'il faudrait vraisemblablement avoir dans une base lexicale. Nous partageons ce point de vue, certaines pistes ont été explorées dans (Schwab, 2005) et continuent à l'être actuellement.

Références

E. AGIRRE, O. ANSA, D. MARTINEZ, et E. HOVY. « Enriching WordNet concepts with topic signatures ». Dans les actes de *NAACL worshop on WordNet and Other Lexical Resources : Applications, Extensions and Customizations*, Pittsburg, USA, 2001.

Romaric BESANÇON. « *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de texte* ». Thèse de doctorat, École Polytechnique Fédérale de Lausanne, Laboratoire d'Intelligence Artificielle, 2001.

Christiane FELLBAUM, . *WordNet : An Electronic Lexical Database*. The MIT Press, 1988.

Olivier FERRET et Michael ZOCK. « Enhancing Electronic Dictionaries with an Index Based on Associations ». Dans les actes de *Proceedings of the 21st International Conference on Computational Linguistics*, pp 281–288, 2006. Association for Computational Linguistics.

Sanda HARABAGIU et Joyce Yue CHAI, . *Usage of WordNet in Natural Language Processing Systems*, Université de Montréal, Montréal, Canada, 1998.

Sanda M. HARABAGIU, George Armitage MILLER, et Dan I. MOLDOVAN. « WordNet 2 - A Morphologically and Semantically Enhanced Resource ». Dans les actes de *Workshop SIGLEX'99 : Standardizing Lexical Resources*, pp 1–8, 1999.

Kevin KNIGHT et Steeve LUK. « Building a Large-Scale Knowledge Base for Machine Translation ». Dans les actes de *AAAI'1994 : National Conference on Artificial Intelligence*, 1994.

Mathieu LAFOURCADE et Frédéric GUINAND. « Ants for Natural Language Processing ». *International Journal of Computational Intelligence Research*, 2006. À paraître.

Mathieu LAFOURCADE et Violaine PRINCE. « Mixing Semantic Networks and Conceptual Vectors : the Case of Hyperonymy ». Dans les actes de *ICCI-2003 (2nd IEEE International Conference on Cognitive Informatics)*, pp 121–128, 2003.

Mathieu LAFOURCADE. « Conceptual Vector Learning - Comparing Bootstrapping from a Thesaurus or Induction by Emergence ». Dans les actes de *LREC'2006*, 2006.

LAROUSSE, . *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse, 1992.

Mathieu MANGEOT-LEREBOURS, Gilles SÉRASSET, et Mathieu LAFOURCADE. « Construction collaborative d'une base lexicale multilingue : Le projet Papillon ». *TAL (Traitement Automatique des langues) : Les dictionnaires électroniques*, pp 151–176, 2003.

Igor MEL'ČUK, André CLAS, et Alain POLGUÈRE. *Introduction à la lexicologie explicative et combinatoire*. Duculot, 1995.

Rada MIHALCEA et Dan MOLDOVAN. « eXtended Wordnet : progress report ». Dans les actes de *NAACL 2001 - Workshop on WordNet and Other Lexical Resources*, Pittsburgh, USA, 2001.

Rada MIHALCEA, Paul TARAU, et Elizabeth FIGA. « PageRank on Semantic Networks, with Application toWord Sense Disambiguation ». Dans les actes de *COLING'2004 : 20th International Conference on Computational Linguistics*, pp 1126–1132, 2004.

Ross QUILLIAN. « *Semantic Informatic processing* », Chapitre Semantic memory, pp 227–270. MIT Press, 1968.

François RASTIER. « *L'isotopie sémantique, du mot au texte* ». Thèse de doctorat d'État, Université de Paris-Sorbonne, 1985.

Didier SCHWAB. « *Approche hybride - lexicale et thématique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte.* ». Thèse de doctorat, Université Montpellier 2, 2005.

Didier SCHWAB, Mathieu LAFOURCADE, et Violaine PRINCE. « Vers l'apprentissage automatique, pour et par les vecteurs conceptuels, de fonctions lexicales. L'exemple de l'antonymie ». Dans les actes de *TALN 2002*, volume 1, pp 125–134, 2002.

Mark STEVENSON. « Augmenting Noun Taxonomies by Combining Lexical Similarity Metrics ». Dans les actes de *COLING'2002 : 19th International Conference on Computational Linguistics*, volume 2/2, pp 953–959, 2002.

Michael ZOCK. « Sorry, What Was Your Name Again, Or How to Overcome The Tip-Of-The Tongue with the help of a computer ? ». Dans les actes de *SemaNet'02 : Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

Alignements monolingues avec déplacements

Julien BOURDAILLET, Jean-Gabriel GANASCIA
Laboratoire d'Informatique de Paris 6
Université Pierre et Marie Curie, 104 Quai Kennedy, 75016 Paris
{julien.bourdaillet,jean-gabriel.ganascia}@lip6.fr

Résumé. Ce travail présente une application d'alignement monolingue qui répond à une problématique posée par la *critique génétique textuelle*, une école d'études littéraires qui s'intéresse à la genèse textuelle en comparant les différentes versions d'une oeuvre. Ceci nécessite l'identification des déplacements, cependant, le problème devient ainsi NP-complet. Notre algorithme heuristique est basé sur la reconnaissance des homologies entre séquences de caractères. Nous présentons une validation expérimentale et montrons que notre logiciel obtient de bons résultats ; il permet notamment l'alignement de livres entiers.

Abstract. This paper presents a monolingual alignment application that addresses a problem which occurs in *textual genetic criticism*, a humanities discipline of literary studies which compares texts' versions to understand texts' genesis. It requires the move detection, but this characteristic makes the problem NP-complete. Our heuristic algorithm is based on pattern matching in character sequences. We present an experimental validation where we show that our application obtains good results ; in particular it enables whole book alignment.

Mots-clés : alignement monolingue, distance d'édition avec déplacements, critique génétique textuelle.

Keywords: monolingual alignment, edit distance with moves, textual genetic criticism.

1 Introduction

L'alignement textuel monolingue consiste à comparer deux textes plus ou moins proches afin d'identifier leurs similitudes et leurs dissemblances ; ou plus précisément, à rechercher les parties communes à ces deux textes et les parties propres à chaque texte. Les premiers travaux d'alignements automatique peuvent être attribués à (Levenshtein, 1966) qui a introduit la distance d'édition : le nombre minimum d'opérations d'édition (insertions, suppressions et remplacements) permettant de transformer un texte en un autre. Par la suite, cette approche considérant les textes comme deux séquences de caractères a été beaucoup étudiée en informatique théorique, voir (Bergroth *et al.*, 2000) pour une synthèse récente, et appliquée où des programmes de comparaison de code source comme *Diff* ont été développés.

Ces méthodes d'alignement de code source, à savoir des langages formels et structurés, ont ensuite été naturellement adaptées pour comparer les textes en langage naturel. Dans les langages formels, on a généralement une seule instruction par ligne ; ainsi entre deux versions d'un fichier, il est relativement simple d'identifier les modifications. Par contre, dans les textes

en langage naturel, une unité entre ligne et phrase n'a pas de raison d'être, si l'on s'en tient au texte et que l'on omet les questions de mise en page liées au support. Et il se trouve en effet que les logiciels d'alignement existants présentent de mauvais résultats pour les textes en langage naturel, comme nous l'avons montré dans (Bourdaillet & Ganascia, 2006).

En Traduction Automatique, il existe une littérature importante sur l'alignement bilingue de textes qui sont généralement la traduction de l'un dans l'autre (bitexte). Ces alignements sont relatifs à des structures de haut niveau, à savoir paragraphes, phrases et plus difficilement mots (Chiao *et al.*, 2006). Nous présentons ici un algorithme d'alignement monolingue au niveau des caractères, entre textes pouvant être très différents l'un de l'autre puisqu'ils peuvent comporter des insertions, suppressions, remplacements et même déplacements. Notre algorithme est plus proche de ceux utilisés en biologie moléculaire tels que (Bray *et al.*, 2003), mais néanmoins il induit un alignement aux niveaux supérieurs.

C'est l'étude des processus de réécriture, dans le cadre d'un travail commun avec l'Institut des Textes et Manuscrits Modernes (ITEM), qui nous a amenée à étudier l'alignement monolingue. C'est dans ce laboratoire qu'est née la *critique génétique textuelle* (de Biasi, 2000), une école d'études littéraires étudiant la genèse des oeuvres littéraires à travers les différents états d'un texte laissés par un écrivain. Ces différentes versions, c'est-à-dire les brouillons successifs, sont annotées par l'auteur qui corrige une faute d'orthographe, affine son vocabulaire ou encore soigne son style en déplaçant un terme. D'un point de vue computationnel, les trois opérateurs classiques de la distance d'édition ne sont pas suffisants pour caractériser ces réécritures ; il est nécessaire d'introduire un opérateur de déplacement d'un bloc de caractères d'une position dans le premier texte vers une position différente dans le second. Cette modélisation correspond à la notion de *distance d'édition avec déplacements*.

Les généticiens du texte ont redécouvert empiriquement cette notion, mais celle-ci avait été introduite auparavant en informatique par (Tichy, 1984). (Lopresti & Tomkins, 1997) ont étendu la notion en introduisant plusieurs modèles de distance d'édition par blocs. (Shapira & Storer, 2002) ont prouvé que le calcul de la distance d'édition avec déplacements entre deux textes est un problème NP-complet ; il n'existe donc pas actuellement d'algorithme le résolvant en un temps polynomial et ils ont proposé un algorithme heuristique glouton pour ce calcul.

L'automatisation de ce travail de comparaison textuelle nécessaire à la critique génétique s'avère donc être un problème difficile. Dans la section 2 nous présentons l'algorithme de notre logiciel, appelé MEDITE[1], traitant ce problème. Dans un précédent travail, nous avions montré son utilité pour la critique génétique (Ganascia & Bourdaillet, 2006). Nous montrons ici que l'algorithme glouton de (Shapira & Storer, 2002) ne permet pas de modéliser correctement ce problème et que MEDITE supporte maintenant le passage à l'échelle en permettant d'aligner différentes versions d'un livre entier (section 3).

L'alignement d'ouvrages complets est une problématique récente née de l'essor des projets de numérisation de livres à grande échelle, comme le "Million Book Project" ou celui de Google (Feng & Manmatha, 2006). De plus, la taille des textes rapproche ce problème de l'alignement des séquences d'acides nucléiques en bioinformatique (Gusfield, 1997). Néanmoins la prise en compte des déplacements n'a pas ou peu été traitée dans ces deux domaines.

Nous pouvons maintenant formuler le problème de manière plus précise. Il consiste à aligner deux textes en langage naturel A et B. Ceux-ci peuvent être vus comme des séquences de caractères de tailles respectives m et n, telles que $A = a_1, a_2, ..., a_m = [a_i]_{1 \leq i \leq m}$ et $B =$

[1] librement téléchargeable en ligne : `http://www-poleia.lip6.fr/~ganascia/medite`

$b_1, b_2, ..., b_n = [b_j]_{1 \leq j \leq n}$ et définies sur un alphabet Σ de taille finie.

Nous définissons la notion de paire de blocs (ou *bi-bloc*) par un tuplet (p, l_A, q, l_B) avec $-1 \leq p \leq |A| = m, 0 \leq l_A \leq m$ et $-1 \leq q \leq |B| = n, 0 \leq l_B \leq n$. Cela signifie qu'une sous-chaine $A[p..p + l_A - 1]$ de la première séquence est en relation avec une sous-chaine $B[q..q + l_B - 1]$ de la seconde séquence.

Finalement, nous définissons un *alignement* $\mathcal{A}(A, B)$ entre deux séquences A et B comme un tuplet tel que $\mathcal{A}(A, B) = (INV, SUP, INS, REMP, DEP)$ avec $INV, SUP, INS, REMP$ et DEP les ensembles de bi-blocs respectivement invariants, supprimés, insérés, remplacés et déplacés constituant cet alignement. Ainsi, le type de relation entre les sous-chaines constituant un bi-bloc est défini par l'ensemble auquel le bi-bloc appartient dans $\mathcal{A}(A, B)$. Les invariants, remplacements et déplacements sont des appariements de blocs effectivement présents dans A et B, alors que les suppressions et insertions sont des pseudo-appariements avec un bloc vide. Pour ce faire, un bloc ayant p ou q égal à -1 représente respectivement une insertion ou une suppression ; dans ce cas l_A ou l_B valent respectivement 0, ce qui correspond à un bloc vide.

2 Algorithme

Notre algorithme se décompose en cinq étapes. La première étape est un pré-traitement qui permet d'établir des classes d'équivalence entre caractères. La seconde étape identifie les blocs de caractères répétés entre A et B. La troisième étape aligne ces blocs répétés afin de déterminer lesquels sont invariants et lesquels sont déplacés. La quatrième étape consiste à répéter les étapes 2 et 3 sur les sous-séquences situées entre les blocs alignés lors de l'étape 3. La dernière étape est la déduction des insertions, suppressions et déplacements. La figure 1 présente cet algorithme.

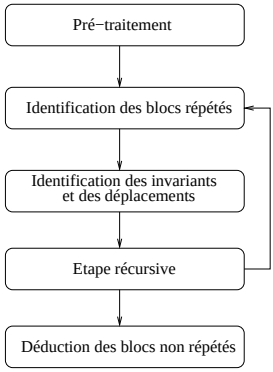


FIG. 1 – Algorithme de MEDITE

2.1 Pré-traitement

Un pré-traitement optionnel peut être appliqué aux séquences A et B. En langage naturel, il existe des caractères que l'on peut vouloir considérer comme équivalents : les caractères identiques mais avec une casse différente, par exemple "J" et "j" ; les caractères avec ou sans signe diacritique, par exemple "ç" et "c" ; ou encore les séparateurs, par exemple " ?" et " !". Pour cela les majuscules sont converties en minuscules, les caractères avec diacritique en leur équivalent sans diacritique, et tous les signes de ponctuation en un même signe, à savoir le caractère point. Ce pré-traitement peut être appliqué sur A et B en un temps linéaire. Ainsi, lors des étapes ultérieures de l'algorithme, des blocs identiques modulo les classes d'équivalence pourront être

appariés même si ces blocs sont différents dans A et B, et ceci pour un coût computationnel faible.

Pour illustrer l'algorithme, nous utilisons l'exemple suivant où nous cherchons à aligner ces deux courts textes : "Ce matin le chat observa de petits oiseaux dans les arbres." et "Le chat était en train d'observer des oiseaux dans les petits arbres ce matin. Il observa les oiseaux pendant deux heures." Après pré-traitement, les phrases deviennent : "ce.matin.le.chat.observa.de.petits. oiseaux.dans.les.arbres." et "le.chat.était.en.train.d.observer.des.oiseaux.dans.les.petits.arbres. ce.matin.il.observa.les.oiseaux.pendant.deux.heures."

2.2 Identification des blocs répétés

L'identification des blocs de caractères répétés, c'est-à-dire présents dans les deux textes, est effectuée en construisant un arbre des suffixes généralisé entre A et B (Ukkonen, 1995). Cette structure de données permet en effet d'identifier l'ensemble des blocs répétés entre A et B en un temps linéaire. Toutefois la taille de cet ensemble de blocs est exponentielle et seul un sous-ensemble est intéressant, celui des appariements exacts super-maximaux (Gusfield, 1997).

Un bi-bloc (p, l_A, q, l_B) est un appariement exact super-maximal si et seulement si :
– $A[p..p + l_A - 1] = B[q..q + l_A - 1]$ (appariement exact) ;
– $A[p - 1] \neq B[q - 1]$ et $A[p + l_A] \neq B[q + l_B]$ (maximalité) ;
– et ni $A[p..p + l_A - 1]$ ni $B[q..q + l_B - 1]$ ne sont inclus dans un autre appariement exact maximal (super-maximalité).

Cette définition n'empêche pas les chevauchements entre appariements exacts super-maximaux (bien que les inclusions le soient). Ces chevauchements peuvent être résolus heuristiquement en les scindant sur les séparateurs, en effet il est préférable d'avoir des coupures entre les mots plutôt qu'à l'intérieur dans les séquences en langage naturel. Le résultat de cette seconde étape se présente sous la forme de deux listes A' et B' de blocs (des séquences A et B) qui ont été identifiés comme faisant partie de l'ensemble des bi-blocs super-maximaux et non-chevauchants.

Dans l'exemple, avant la résolution des chevauchements, les blocs super-maximaux "s.arbres." chevauchent les blocs "oiseaux.dans.les." et "petits." ; la césure sur le séparateur permet de résoudre le conflit. Finalement, après la seconde étape, les blocs super-maximaux non-chevauchants suivants sont identifiés : "ce.matin. le.chat. observa. de. petits. oiseaux.dans.les. arbres." et "le.chat. était.en.train.d.observer.des. oiseaux.dans.les. petits. arbres. ce. matin. il. observa. les.oiseaux. pendant.deux.heures.". Le mot "oiseaux" est répété trois fois mais n'apparait pas dans la liste des blocs super-maximaux car les deux premières occurrences sont incluses dans des blocs plus longs qui eux sont super-maximaux.

2.3 Identification des blocs invariants et déplacés

Les blocs invariants sont ceux qui apparaissent à la même position dans A et B, et les déplacés ceux dont la position a changé. Or chacun des blocs super-maximaux identifiés lors de l'étape précédente peut être soit un bloc invariant, soit un bloc déplacé. En effet, lorsque deux blocs sont permutés, on peut établir que l'un est invariant et l'autre déplacé ou vice-versa, et il n'existe pas de variable permettant de prendre la bonne décision de façon certaine. Néanmoins nous pouvons utiliser le critère heuristique suivant : entre une telle paire de blocs, le plus long sera considéré

comme invariant et l'autre comme déplacé.

La méthode exhaustive permettant de prendre l'ensemble de ces décisions pour tous les blocs consiste à parcourir l'espace des alignements possibles afin de trouver l'optimum suivant une certaine fonction de coût. Or la taille de cet espace est combinatoire, ce qui rend cette recherche peu opérationnelle. C'est pourquoi nous utilisons l'algorithme de type A* suivant.

Les alignements possibles sont évalués à l'aide d'une fonction de coût c ; l'objectif est de trouver un alignement de coût minimal. Ceci est équivalent à un problème de plus court chemin dans un graphe où l'état final correspond à l'alignement de coût minimal et l'état initial à l'état où aucune décision n'a encore été prise. A chaque étape de l'algorithme, une décision est prise en choisissant de désigner l'appariement d'un bloc A'_i avec un bloc B'_j comme étant un bi-bloc invariant. Ce choix est conduit grâce à la fonction de coût c qui estime le coût de l'alignement final induit par ce choix. Lorsque l'état final est atteint, c'est-à-dire lorsque l'on ne peut plus choisir de bloc invariant, tous les blocs qui n'ont pas été choisis durant le parcours sont considérés comme des blocs déplacés. Afin d'atteindre l'état final, c doit être admissible, c'est-à-dire ne jamais surestimer le coût de l'alignement ; nous détaillons ci-dessous pourquoi c est admissible.

L'évaluation du coût de l'alignement induit par le choix de l'appariement de A'_i et B'_j est calculé par la fonction $c(i,j)$. Celle-ci décompose ce coût en un coût $g(i,j)$ de l'alignement effectué lors des étapes précédentes, et une estimation heuristique $h(i,j)$ du coût de l'alignement qu'il reste à effectuer durant les étapes ultérieures, tel que $c(i,j) = g(i,j) + h(i,j)$. Ces coûts sont calculés de la façon suivante :

- $NA(i,j) = non-appariés(A'[1..i-1], B'[1..j-1])$ est l'ensemble des blocs non appariés durant les étapes précédentes, ceux qui n'ont pas été choisis comme invariants et seront considérés comme déplacés.
- $g(i,j) = \sum_{b \in NA(i,j)} |b|$ est la somme de la taille des blocs précédemment non choisis comme étant invariants, c'est-à-dire que seuls les déplacements vont pénaliser le coût d'un alignement.
- $DS(i,j) = A'[i+1..|A'|] \ominus B'[j+1..|B'|]$ est la différence symétrique des deux ensembles de blocs à aligner durant les étapes suivantes. Ces blocs sont présents soit uniquement dans $A'[i+1..|A'|]$ soit uniquement dans $B'[j+1..|B'|]$, il ne sera donc pas possible de les apparier ultérieurement.
- $h(i,j) = \sum_{b \in DS(i,j)} |b|$ est la somme de la taille des blocs de $DS(i,j)$. $h(i,j)$ est la borne inférieure du coût des blocs restant à aligner. h ne surestime jamais le coût de l'alignement, c'est pourquoi c est admissible et A* trouve l'alignement optimal au sens de notre critère.

Ce calcul permet de rechercher un alignement optimal au sens de la maximisation de la taille des blocs invariants et de la minimisation de la taille des blocs déplacés.

Dans notre exemple, après cette étape, les blocs encadrés en gras désignent les invariants et les autres blocs encadrés les déplacements : "ce.matin. **le.chat.** observa. de. petits. **oiseaux.dans.les.** **arbres.**" et "**le.chat.** était.en.train.d.observer.des. **oiseaux.dans.les.** petits. **arbres.** ce. matin. il. observa. les.oiseaux. pendant.deux.heures.".

2.4 Recherche récursive d'appariements

Lors de cette quatrième étape, on considère chaque sous-chaine de A et B située entre deux bi-blocs invariants. Ces sous-chaines sont examinées à nouveau par les étapes 2 et 3 afin de découvrir d'éventuels nouveaux bi-blocs invariants, auquel cas ceux-ci sont ensuite inclus dans

l'alignement principal. Cette étape récursive permet de répondre aux *effets de masquage* qui se produisent lorsque les séquences A et B comportent un nombre important de sous-chaines répétées : dans ces cas là, les algorithmes classiques d'alignement omettent des appariements importants qui sont masqués par des appariements moins importants (Ganascia & Bourdaillet, 2006). De tels phénomènes ont également été identifiés dans les séquences d'acides nucléiques en biologie moléculaire (Arslan *et al.*, 2001).

Dans notre exemple, entre les bi-blocs invariants "le.chat." et "oiseaux.dans.les." se trouvent les sous-chaines "observa.de.petits." et "était.en.train.d.observer.des.". L'étape récursive va permettre d'identifier le bi-bloc "observ" comme invariant, ce qui donne l'alignement final suivant : "ce.matin. **le.chat.** **observ**a.de. petits. **oiseaux.dans.les.** **arbres.**" et "**le.chat.** était.en.train.d.**observ**er.des. **oiseaux.dans.les.** petits. **arbres.** ce. matin. il.observa.les. oiseaux.pendant.deux.heures.". Ainsi, un bi-bloc déplacé a été perdu ("observa.") mais un bi-bloc invariant a été gagné ("observ") ; ceci permet de favoriser les appariements locaux au détriment des appariements longue-distance et les invariants plutôt que les déplacements.

2.5 Déduction des autres types de blocs

Les insertions, suppressions et remplacements peuvent finalement être déduits des étapes précédentes. En effet, les suppressions sont les blocs non répétés et présents uniquement dans A, et les insertions ceux présents uniquement dans B.

L'identification des remplacements se fait de manière heuristique : lorsqu'entre deux bi-blocs invariants se trouve un bloc supprimé s dans A et un bloc inséré i dans B et que le ratio entre leur taille $|s|/|i|$ atteint un certain seuil t, alors ces blocs sont retirés des ensembles SUP et INS (cf. section 1), et appariés en un bi-bloc r placé dans $REMP$, signifiant ainsi que le bloc dans A a été remplacé par le bloc dans B. Le seuil t est fixé par défaut à $0,5$.

Dans l'exemple, le bi-bloc constitué des chaines "a.de." et "er.des." sera considéré comme un remplacement et les deux autres blocs non encadrés de la seconde séquence comme des insertions.

Finalement, un post-traitement permet de retrouver les positions des blocs dans les séquences originales (c'est-à-dire sans les classes d'équivalence).

3 Validation expérimentale

3.1 Application à la critique génétique textuelle

Cette expérience consiste à aligner deux versions d'un même texte et évaluer l'alignement résultant. Pour ce faire, nous allons comparer les résultats de MEDITE à ceux de GREEDY qui est un algorithme glouton de calcul de la distance d'édition avec déplacements (Shapira & Storer, 2002). Ce dernier sélectionne à chaque itération le plus grand appariement qu'il considère comme un déplacement et finalement calcule une distance d'édition classique par programmation dynamique.

L'évaluation des alignements résultants se fait en calculant les fonctions de score suivantes à partir d'un alignement $\mathcal{A}(A, B)$:

– Nous définissons au préalable une fonction $somme(S) = \sum_{(p,l_A,q,l_B)\in S} l_A + l_B$ qui somme la taille de tous les bi-blocs d'un ensemble de bi-blocs S.
– On cherche à maximiser la somme de la taille des blocs invariants et à minimiser la somme de la taille des autres types de bloc, d'où la fonction :

$$x = \left(1 + \frac{somme(INV) - \sum_{s\in S} somme(s)}{|A| + |B|}\right)/2 \quad \text{avec } S = \{SUP, INS, REMP, DEP\} \tag{1}$$

– On cherche à maximiser la taille moyenne des blocs afin d'éviter la fragmentation de l'alignement :

$$y = \left(\sum_{s\in S}\left(\frac{somme(s)}{|s|}\right) / \max(s)\right)/5 \quad \text{avec } S = \{INV, SUP, INS, REMP, DEP\}, \tag{2}$$
$|s|$ le nombre de blocs dans s
et $\max(s)$ la taille du plus grand bloc de s

– On cherche à maximiser le ratio des déplacements par rapport aux autres blocs non-invariants et le ratio des remplacements par rapport aux autres blocs non-invariants (sauf les déplacements) :

$$z = \left(\frac{somme(MOV)}{somme(S_1)} + \frac{somme(REMP)}{somme(S_2)}\right)/2 \quad \text{avec } S_1 = \{SUP, INS, REMP, DEP\} \text{ et } S_2 = \{SUP, INS, REMP\} \tag{3}$$

– Finalement, ceci nous permet de définir une fonction de similarité globale combinant les équations précédentes ; les pondérations sont fixées arbitrairement mais reflètent les priorités accordées aux différentes fonctions :

$$sim = 0.5x + 0.35y + 0.15z \tag{4}$$

Les termes de normalisation rendent ces équations un peu chargées, mais les idées sous-jacentes sont très simples.

Les textes à aligner sont les suivants : deux versions d'un poème d'Andrée Chedid "La Robe Noire" de 2 Ko, nommé A ci-dessous ; un cahier d'expérience de Claude Bernard et une synthèse académique de ce cahier (7.5 Ko, B) ; un sous-ensemble de la partie française du Hansard et la traduction en français de la partie correspondante anglaise[2] (20 Ko, C) ; et deux versions d'un texte de Louis Althusser "Freud et Lacan" (50 Ko, D). Le tableau 1 présente les résultats de ces alignements.

On peut constater que MEDITE obtient de meilleurs résultats pour tous les textes et critères (sauf pour z sur B et C). Le critère x signifie que MEDITE trouve plus de blocs invariants que GREEDY ; y signifie que les blocs alignés sont plus longs ; et z qu'on favorise les déplacements au détriment des autres types de blocs non-invariants et les remplacements au détriment des insertions et suppressions, en effet ces blocs apportent plus d'informations. On remarquera aussi les différences considérables en temps de calcul.

[2] corpus pré-traité et mis à disposition par le RALI

Algorithme	GREEDY				MEDITE			
Texte	A	B	C	D	A	B	C	D
x	0.3654	0.2657	0.4106	0.7835	0.4934	0.2697	0.4936	0.9223
y	0.1161	0.0793	0.0784	0.1397	0.3331	0.2488	0.1951	0.2318
z	0.1971	0.2340	0.4096	0.1653	0.2003	0.1676	0.2937	0.2587
sim	0.2529	0.1957	0.2942	0.4655	0.3933	0.2471	0.3591	0.5811
Temps	0mn 18s	12mn 5s	1h 1mn	29mn 3s	0mn 1s	0mn 2s	0mn 6s	0mn 2s

TAB. 1 – Alignements avec GREEDY et MEDITE

3.2 Alignement de données synthétiques

Le but de cette seconde expérience est d'évaluer la qualité des alignements de MEDITE sur des données synthétiques où il existe un alignement de référence. Etant donné un texte et un générateur de bruit, un second texte est généré en altérant le premier. L'alignement entre les deux textes est enregistré durant le processus d'altération ; il est alors possible d'évaluer la qualité d'un aligneur en comparant ses résultats avec l'alignement de référence.

Premier générateur de bruit Le générateur de bruit permet de générer un second texte à partir de l'original de la façon suivante. Des ratios d'insertions, suppressions et remplacements sont fixés avant de commencer. Des blocs de caractères sont alors insérés dans le second texte, supprimés dans l'original et remplacés entre les deux textes, de façon répétée jusqu'à ce que les ratios soient atteints. Les positions des modifications sont choisies aléatoirement sur toute la longueur des textes (le chevauchement d'opérations n'est pas permis). La taille des blocs est choisie aléatoirement entre 1 et 25 caractères. Durant ce processus, les positions des modifications sont enregistrées, ce qui permet d'obtenir un alignement de référence.

Pour cette expérience, nous avons choisi un texte de 520 Ko comme texte original, soit la taille d'un livre d'environ 350 pages. Cinq textes synthétiques différents ont été générés et alignés chacun avec l'original via MEDITE, puis les scores de précision calculés. Deux séries de tests avec différents ratios de modifications ont été menées : dans la première il y a 5% d'insertions, 5 % de suppressions et 5 % de remplacements, ce qui signifie que les textes altérés présentent 15 % de différences avec l'original ; dans la seconde série, les ratios sont portés à 10 %, ce qui signifie qu'il y 30 % de différences entre les textes.

Pour chacun des quatre types de caractères (invariants, insertions, suppressions et remplacements) le taux de précision est défini comme le nombre de caractères correctement alignés / le nombre total de ces caractères. Les précisions moyennes sur les cinq alignements sont alors calculées. Pour la précision pondérée, les précisions de chaque type sont pondérées par leurs poids respectifs dans les textes ; par exemple pour la première série de tests on aura $Prec.pondérée = 0.85 * Prec.INV + 0.05 * Prec.INS + 0.05 * Prec.SUP + 0.05 * Prec.REMP$. Les deux premières colonnes du tableau 2 présentent les résultats de cette expérience.

On peut constater que les précisions moyennes sont bonnes, en particulier les précisions pondérées, et que les temps de calcul sont raisonnables ; il serait inenvisageable de traiter ces textes avec GREEDY. L'expérience a été réalisée sur un Pentium 4, 2.4 GHz avec 1 Go de RAM. MEDITE est implémenté en Python, un langage de haut niveau, bon pour le prototypage mais

Générateur de bruit	sans déplacements		avec déplacements	
Ratio de modifications	5 %	10 %	5 %	10 %
Précision moyenne	94.48 %	89.27 %	86.56 %	78.36 %
Précision pondérée moyenne	98.16 %	94.0 %	95.19 %	86.18 %
Temps moyen	11 mn 5 s	27 mn 53 s	27 mn 8 s	77 mn 17 s

TAB. 2 – Alignement de données synthétiques avec MEDITE

lent. Une implémentation en C permettrait de gagner considérablement en vitesse d'exécution. Néanmoins, le goulot d'étranglement de notre algorithme reste le calcul des différences symétriques entre listes de blocs (voir section 2.4), qui est quadratique par rapport à la taille de ces listes.

Générateur de bruit avec déplacements Ce second générateur de bruit est similaire au premier mais en plus des déplacements seront générés entre texte original et texte altéré. Ainsi, des blocs de caractères sont déplacés d'une position dans le texte original vers une seconde dans le texte altéré. Les ratios de modifications sont toujours fixés à 5 et 10 % par opérations, ce qui donne des textes avec 20 et 40 % de différences respectivement. Deux séries de tests sont à nouveau conduites et les moyennes des résultats présentées dans les deux dernières colonnes du tableau 2.

On peut constater que les précisions moyennes décroissent significativement mais que les précisions pondérées conservent de meilleurs scores. Il faut toutefois garder à l'esprit que les ratios de différences sont de 20 et 40 % contre 15 et 30 % avec le premier générateur de bruit. De plus, la différence entre les précisions pondérées et non-pondérées indiquent que les blocs invariants ont un meilleur taux de classification. Ceci est confirmé dans le tableau 3 qui présente la moyenne des matrices de confusion : les blocs de référence sont en lignes et ceux trouvés par MEDITE en colonnes. Les erreurs les plus importantes proviennent des déplacements qui sont identifiés comme insertions et suppressions ; or un déplacement peut être considéré comme une suppression suivie d'une insertion. De même les insertions et suppressions sont confondues avec des remplacements ; or les remplacements peuvent aussi être considérés comme une suppression suivie d'une insertion. Ceci implique un problème de décision : notre modèle de décision est très simple et pourrait être amélioré ; néanmoins les résultats présentés ont le mérite d'être consistants.

Ratio de modifications	5 %					10 %				
Type de bloc	INV	INS	SUP	REMP	DEP	INV	INS	SUP	REMP	DEP
Invariants	**98.07**	0.56	0.52	0.76	0.08	**94.0**	1.75	1.6	2.32	0.32
Insertions	0.21	**92.16**	0	7.55	0.07	0.21	**85.1**	0	14.52	0.17
Suppressions	0.14	0	**87.76**	9.25	1.76	1.46	0	**76.77**	17.78	3.99
Remplacements	0.70	5.40	4.65	**88.40**	0.84	0.72	11.32	9.43	**76.34**	2.18
Déplacements	1.43	13.70	14.05	4.38	**66.43**	1.47	15.16	15.55	8.25	**59.56**

TAB. 3 – Moyenne des matrices de confusion (en %) pour le générateur de bruit avec déplacements

4 Conclusion

Nous avons présenté MEDITE, un aligneur monolingue détectant les déplacements entre deux textes. Nous traitons ce problème d'alignement difficile par un algorithme heuristique de recherche d'homologies dans les séquences. Notre validation expérimentale montre que MEDITE présente de bons résultats et qu'il est capable d'aligner des livres entiers en un temps raisonnable, tout en identifiant les déplacements.

MEDITE est maintenant utilisé par les généticiens du texte pour aligner différentes versions de livres entiers. Ce travail fastidieux nécessiterait plusieurs mois, voire plusieurs années de travail sans l'usage de la machine. Nous projetons maintenant de l'utiliser pour établir des éditions électroniques d'ouvrages en intégrant directement le logiciel dans le support électronique.

Références

ARSLAN A. N., EGECIOGLU O. & PEVZNER P. A. (2001). A new approach to sequence comparison : normalized sequence alignment. *Bioinformatics*, **17**(4), 327–337.

BERGROTH L., HAKONEN H. & RAITA T. (2000). A Survey of Longest Common Subsequence Algorithms. In *SPIRE '00 : Proceedings of the Seventh International Symposium on String Processing Information Retrieval*.

BOURDAILLET J. & GANASCIA J.-G. (2006). MEDITE : A unilingual textual aligner. *In Proceedings of FinTAL, 5th International Conference on Natural Language Processing, Lecture Notes in Artificial Intelligence*, **4139**, 458–469.

BRAY N., DUBCHAK I. & PACHTER L. (2003). AVID : A Global Alignment Program. *Genome Res.*, **13**(1), 97–102.

CHIAO Y.-C., KRAIF O., LAURENT D., NGUYEN T. M. H., SEMMAR N., STUCK F., VÉRONIS J. & ZAGHOUANI W. (2006). Evaluation of multilingual text alignment systems : the ARCADE II project. *Proceedings of the LREC 2006 Conference*.

DE BIASI P.-M. (2000). *La Génétique des Textes*. Nathan Université.

FENG S. & MANMATHA R. (2006). A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *JCDL'06*, p. 109–118 : ACM Press.

GANASCIA J.-G. & BOURDAILLET J. (2006). Alignements unilingues avec MEDITE. In *8^{émes} Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2006)*.

GUSFIELD D. (1997). *Algorithms on Strings, Trees and Sequences : Computer Science and Computer Biology*. Cambridge University Press.

LEVENSHTEIN V. (1966). Binary codes capable of correcting deletions, insertions and reversal. *Cybernetics and Control Theory*, **10**(8), 707–710.

LOPRESTI D. P. & TOMKINS A. (1997). Block Edit Models for Approximate String Matching. *Theoretical Computer Science*, **181**(1), 159–179.

SHAPIRA D. & STORER J. A. (2002). Edit Distance with Move Operations. In *CPM*, volume 2373 of *Lecture Notes in Computer Science*, p. 85–98 : Springer.

TICHY W. F. (1984). The String-to-String Correction Problem with Block Moves. *ACM Trans. Comput. Syst.*, **2**(4), 309–321.

UKKONEN E. (1995). On-Line Construction of Suffix Trees. *Algorithmica*, **14**(3), 249–260.

Session

Syntaxe

Confondre le coupable : corrections d'un lexique suggérées par une grammaire

Lionel NICOLAS[1], Jacques FARRÉ[1], Éric VILLEMONTE DE LA CLERGERIE[2]
[1] Laboratoire I3S, Université de Nice-Sophia Antipolis, CNRS
2000 route des Lucioles, B.P. 121, 06903 Sophia Antipolis Cedex, France
[2] Projet ATOLL - INRIA
Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France
{lnicolas,jf}@i3s.unice.fr,
Eric.De_La_Clergerie@inria.fr

Résumé. Le succès de l'analyse syntaxique d'une phrase dépend de la qualité de la grammaire sous-jacente mais aussi de celle du lexique utilisé. Une première étape dans l'amélioration des lexiques consiste à identifier les entrées lexicales potentiellement erronées, par exemple en utilisant des techniques de fouilles d'erreurs sur corpus (Sagot & Villemonte de La Clergerie, 2006). Nous explorons ici l'étape suivante : la suggestion de corrections pour les entrées identifiées. Cet objectif est atteint au travers de réanalyses des phrases rejetées à l'étape précédente, après modification des informations portées par les entrées suspectées. Un calcul statistique sur les nouveaux résultats permet ensuite de mettre en valeur les corrections les plus pertinentes.

Abstract. Successful parsing depends on the quality of the underlying grammar but also on the quality of the lexicon. A first step towards the improvement of lexica consists in identifying potentially erroneous lexical entries, for instance by using error mining techniques on corpora (Sagot & Villemonte de La Clergerie, 2006). we explores the next step, namely the suggestion of corrections for those entries. This is achieved by parsing the sentences rejected at the previous step anew, after modifying the information carried by the suspected entries. Afterwards, a statistical computation on the parsing results exhibits the most relevant corrections.

Mots-clés : analyse syntaxique, lexique, apprentissage, correction .

Keywords: parsing, lexicon, machine learning, correction .

1 Introduction

L'analyse syntaxique d'une langue repose sur l'utilisation de ressources linguistiques les plus précises et correctes possibles. Obtenir des ressources possédant une si large couverture est une tâche ardue de longue haleine qu'il est souhaitable d'alléger par le biais de techniques qui en automatisent l'élaboration et la correction. Nous présentons ici une technique de génération automatique de suggestions de corrections pour les entrées potentiellement erronées d'un lexique.

Nous nous intéressons aux moyens de réduire l'inexactitude et l'incomplétude d'un lexique à partir d'un recensement de formes lexicales suspectées d'être incorrectement ou seulement

partiellement décrites dans un lexique. Nous nous situons ainsi dans le prolongement direct de la technique de fouille d'erreurs sur des corpus de grande taille originalement proposée par (van Noord, 2004), et améliorée par (Sagot & Villemonte de La Clergerie, 2006). La pertinence de cette dernière s'observe notamment à travers nos résultats.

Cette technique repose sur l'idée suivante : étant donné un large corpus de phrases attestées, plus une forme (et indirectement les lemmes associés) apparaît ou n'apparaît pas dans des phrases dont les analyses échouent, plus nous avons des raisons de douter ou de ne pas douter des entrées lexicales qui lui sont associées. Cependant le contexte des formes importe : une forme est d'autant plus suspecte qu'elle apparaît dans des phrases non analysables mais en co-occurrence avec des formes qui tendent à apparaître dans des phrases analysables.

L'implémentation de la technique de fouilles d'erreurs nous a fourni une liste de 5344 formes suspectes avec, pour chaque forme f, un taux de suspicion et une liste de phrases non analysables (56089 au total) où f est suspectée d'être à l'origine de l'échec des analyses. Si une forme est effectivement responsable de ces échecs, et non la grammaire[1], c'est donc que les informations lexicales qui lui sont associées sont incomplètes ou inexactes (voir inexistantes).

En relâchant les contraintes sur les informations portées par une forme suspecte ou en les modifiant (notamment la catégorie syntaxique), de nouvelles analyses des phrases associées vont aboutir. Les représentations des phrases alors produites représentent les conditions dans lesquelles l'analyse a réussi, c.a.d. les informations sur la forme suspecte rendant possible l'analyse. En examinant ces informations sur un ensemble de phrases, il est alors possible de dégager des hypothèses de correction utiles.

La technique présentée est indépendante du langage étudié.

Travaux relatifs. L'acquisition de connaissances linguistiques depuis des corpus bruts (i.e. non annotés) par le biais de connaissances grammaticales a été initialement étudiée par (Brent, 1993) afin d'identifier les cadres syntaxiques des verbes en anglais. (Horiguchi *et al.*, 1995) utilisent les résultats d'analyse fournis par un système HPSG afin d'acquérir des entrées lexicales de mots japonais inconnus. Enfin, mentionnons la reconstitution d'informations lexicales manquantes en vue d'analyses robustes (Grover & Lascarides, 2001), (Crysmann *et al.*, 2002).

Nous commençons par expliquer comment générer des hypothèses de correction (Sect. 2) et comment les trier (Sect. 3). Nous introduisons ensuite la notion de synchronisation entre un lexique et une grammaire (Sect. 4), juste avant d'exposer les résultats obtenus (Sect. 5) et les développements futurs (Sect. 6).

2 Génération d'hypothèses

Le principal but d'un analyseur syntaxique est de vérifier la validité syntaxique d'une phrase et d'en produire une ou plusieurs représentations. On souhaite en général éviter la surgénération des représentations issues d'une analyse en produisant le moins possible de représentations.

Une phrase est qualifiée d'*ambiguë* pour un analyseur lorsque celui-ci lui associe plusieurs interprétations. Ceci arrive principalement lorsque la phrase est intrinsèquement ambiguë, i.e. d'autres informations (tel que le contexte sémantique) sont nécessaires afin de filtrer les inter-

[1]Nous supposons que les erreurs dues à un traitement incorrect en amont du processus d'analyse proprement dit (segmentation, ponctuation, détection d'entités nommées, ...) ont été identifiées. Les formes erronées et leurs phrases associées qui résulteraient de telles erreurs sont donc exclues de celles qui nous intéressent ici.

prétations, ou lorsque les ressources utilisées (lexique, grammaire ...) ne sont pas assez restrictives et acceptent un langage plus large.

Afin de rejeter les phrases n'appartenant pas à la langue, on souhaite disposer d'un lexique le plus précis et détaillé possible. En effet, plus une forme lexicale est spécifiée, moins elle se combine avec les autres constituants de la phrase, et par conséquent, moins elle permet d'interprétations incorrectes.

2.1 Causes d'échec d'une analyse

Chaque forme possède, à travers ses lemmes, différentes informations pouvant être regroupées en deux ensembles : d'une part la catégorie syntaxique (nom, verbe, adjectif, ...), d'autre part les informations morphologiques (nombre, genre, personne, temps, mode, ...) et syntaxiques (valence, facultativité des arguments, réflexivité, passivation, ...). L'échec d'une analyse à cause d'une forme est la conséquence d'un problème touchant à au moins un de ces ensembles.

2.1.1 Défaut de catégorisation

Une forme peut être associée à plusieurs lemmes (homonymes) avec des catégories syntaxiques distinctes. Le traitement de telles formes ambiguës au sein d'une phrase se gère par le passage d'un treillis de mots (ou DAG) à l'analyseur syntaxique (Sagot & Boullier, 2005). Une analyse syntaxique réussie valide au moins un chemin possible de lecture dans ce treillis.

Cependant, un lexique peut ne pas recenser tous les homonymes d'une forme et induire ainsi des échecs d'analyse. Par exemple, la forme « fiche » dénote un nom commun et une flexion du verbe « ficher » . S'il n'existe aucun lemme associé de catégorie *nom-commun*, la phrase « Ma fiche contient une erreur. » sera représentée par une seule séquence de catégories *ma/pronom-possessif fiche/verbe contient/verbe une/det erreur/nom-commun*. À moins qu'une production grammaticale n'accepte une telle construction, son analyse devrait aboutir à un échec.

2.1.2 Sur-spécification

En général, on associe aux règles de grammaire des décorations, exprimées sous formes de structures de traits et chargées de compléter les vérifications amorcées par le squelette syntaxique d'une production grammaticale (Abeillé, 1993). Par exemple, un squelette vérifie la présence d'un groupe nominal sujet et d'un verbe dans une phrase là ou les décorations en vérifient l'accord (même personne, nombre, et éventuellement genre).

Comme nous l'avons expliqué, il est souhaitable que les formes lexicales soient les plus spécifiées possible afin de réduire les ambiguïtés. En revanche, si ces dernières sont trop restrictives (autrement dit sur-spécifiées), certaines analyses échouent à cause du mécanisme d'unification des décorations de la grammaire et des restrictions d'utilisation des entrées lexicales.

Il est par exemple très difficile de renseigner un verbe sur l'ensemble de ses emplois possibles, du fait de la polysémie, de la facultativité de certains arguments, de possibles alternations (« acheter qchose » donnant « qchose s'achète »), et de multiples réalisations des arguments (« aimer qchose » , « aimer que + S » , « aimer Sinf »). Il arrive donc que l'on considère comme obligatoires des aspects qui ne sont que facultatifs dans certains cas. Ce constat s'étend aux autres catégories syntaxiques dès lors qu'on leur attache des cadres de catégorisation.

2.2 Réanalyser en sous-spécifiant

Puisque seules les phrases dont l'analyse a échoué sont conservées durant l'étape de fouille d'erreurs, leur taux d'analyse est nul. Si une modification des informations lexicales portées par une forme suspecte f permet d'augmenter sensiblement le taux de réanalyse des phrases qui lui sont associées, il est raisonnable de penser que le problème est bien lié à f. La difficulté est alors de trouver quelles modifications permettent des augmentations sensibles. Plutôt que de tester toutes les combinaisons de modifications possibles, ce qui est exponentiel, nous nous reposons sur la capacité de notre analyseur à pouvoir gérer des formes sous-spécifiées.

Une fois obtenus de nouveaux résultats d'analyse, nous sommes en mesure d'en extraire des hypothèses de correction (voir Sect. 2.2.2).

2.2.1 Génération et utilisation de jokers

Afin de rendre analysables des phrases qui ne l'étaient initialement pas, nous introduisons à la place des formes lexicales suspectes des formes sous-spécifiées appelées *jokers*. Dans l'approche actuelle (qui demande à être affinée), elles ne possèdent qu'une catégorie syntaxique (parmi les catégories « ouvertes » : verbe, nom commun, adjectif ou adverbe). Elles n'ont donc aucune information morphologique ou syntaxique fixe et remplissent toujours les conditions fixées par les décorations des productions grammaticales. Puisque leur utilisation ne soulève aucun conflit lors des analyses (excepté pour la catégorie syntaxique), les substituer à une forme suspecte dans une phrase rejetée favorise la réussite de son analyse. Cependant, cela peut introduire une certaine ambiguïté car il n'y a plus de filtrage au niveau des décorations.

Étant donné que nous ne pouvons savoir *a priori* quel type d'erreur (sur-spécification ou défaut de catégorisation) est responsable des échecs d'analyse, nous considérons les deux simultanément au moment de générer les jokers.

Pour envisager une sur-spécification, nous remplaçons une forme de catégorie X par un joker de même catégorie X. Les caractéristiques permettent alors d'explorer les mêmes productions grammaticales que pour la phrase initiale, sans pour autant être arrêté par les décorations.

Pour faire face à un défaut de catégorisation d'une forme f, nous créons des jokers avec des catégories syntaxiques différentes de celles initialement recensées pour f. En procédant ainsi, les réanalyses exploreront d'autres productions. Ces jokers sont générés à partir des informations fournies par un lemmatiseur (*stemmer*) ou par un tagger probabiliste tel que TREETAGGER (Schmid, 1999).

Nous aurions pu utiliser un joker unique ne possédant même pas de catégorie syntaxique et permettant de couvrir à lui seul l'ensemble des situations décrites ci-dessus. Cependant, un tel joker introduit une très forte ambiguïté, aboutissant soit à un échec des analyses par limite de temps ou de mémoire, soit à la surgénération de représentations pour une phrase. Dans le premier cas, nous ne collectons aucune donnée, dans le second cas, le volume de données est trop important pour être correctement trié et valorisé. Notre approche permet (en grande partie) d'écarter ces problèmes tout en évitant de multiplier le nombre de jokers par forme suspecte.

Nous avons testé une moyenne de 2.05 jokers par forme suspecte (10978 au total), donnant lieu à 117655 nouvelles analyses.

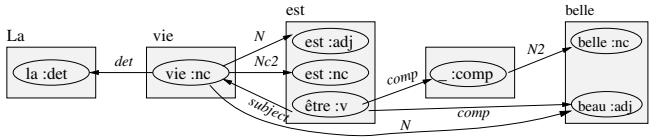


FIG. 1 – Extrait de la représentation graphique d'une forêt partagées de dépendances

2.2.2 Extraction de signatures syntaxiques

Si une forme suspecte a été correctement identifiée, son remplacement par des jokers dans les phrases qui lui sont associées permet à certaines analyses de réussir. Dans les faits, on observe une relation nette entre le taux de succès de l'analyse des phrases modifiées et le taux de suspicion de la forme concernée.

Notre analyseur renvoie l'ensemble des interprétations possibles d'une phrase sous la forme d'une forêt partagée de dépendances (Fig. 1) où les nœuds représentent les lemmes et les arcs les dépendances syntaxiques entre les lemmes. Chaque nœud possède des informations relatives au lemme et à la production grammaticale ancrée (dans le cadre d'une grammaire lexicalisée). Chaque dépendance est caractérisée par un nœud gouverneur source, un nœud gouverné cible, une nature et un label qui dépend de la grammaire. Ce label dénote souvent (mais malheureusement pas toujours) la fonction syntaxique de la cible (sujet, objet, ...). Afin de gérer les ambiguïtés, des informations complémentaires locales au nœud gouverneur lient les lemmes et les dépendances à une ou plusieurs interprétations. Ainsi, la représentation issue de l'analyse de la phrase pour « La vie est belle » (Fig. 1) donne lieu à quatre lectures possibles, du fait (a) de l'ambiguïté de « est » comme verbe à copule, nom commun (en apposition de « vie ») et adjectif ainsi que (b) de l'ambiguïté de « belle » entre adjectif et nom.

Sans aucune information supplémentaire, les deux interprétations comme nom et adjectif de « est » auraient dû être rejetées car introduisant une apposition rare et/ou construisant une phrase sans verbe.

Les forêts contiennent donc les dépendances entrantes et sortantes depuis et vers un joker. Nous appelons désormais *signature syntaxique* l'ensemble de dépendances autour d'un joker dans une interprétation particulière et *groupe de signatures* l'ensemble des signatures syntaxiques possibles extraites des interprétations obtenues par l'analyse réussie d'une phrase.

Ces signatures représentent les conditions dans lesquelles l'analyse a pu aboutir, i.e. les données que la grammaire aurait accepté pour la forme suspecte. Du fait de l'ambiguïté consécutive à l'introduction d'un joker, un analyseur peut produire plusieurs interprétations et donc plusieurs signatures. Parmi ces interprétations, une est plus proche du sens réel de la phrase que les autres. La signature qu'elle contient possède alors les données les plus pertinentes et intéressantes, celles que nous recherchons afin de déterminer les corrections à appliquer au lexique.

3 Identifier les meilleures signatures

En se plaçant au niveau d'un seul groupe de signatures (produit à partir d'une seule phrase), nous sommes incapables de différencier les signatures pertinentes de celles qui ne sont qu'une

conséquence de l'ambiguïté introduite par le joker.

La variabilité de contexte induite par plusieurs groupes de signatures (produits à partir de plusieurs phrases) nous apporte une solution à ce problème. En effet, elle implique la diversification des signatures « parasites » qui contraste avec la stabilité des signatures pertinentes représentant le(s) sens réel(s) de la forme.

Une répétition bien marquée de certaines signatures sur l'ensemble des phrases suggère alors un schéma d'utilisation attendu par la grammaire pour la forme. Afin de pouvoir l'observer, nous valorisons/dévalorisons les signatures par le biais d'un calcul statistique simple en deux étapes.

Première étape : distribution locale des poids entre signatures. L'intérêt que nous portons à un groupe de signatures dépend de sa taille : plus il contient de signatures moins il présente d'intérêt. En effet, il est vraisemblable que plusieurs squelettes syntaxiques « permissifs » lui correspondent, à l'image de ceux permettant les diverses interprétations illustrées par la figure 1. Pour chaque groupe g, nous calculons donc un poids $P = c^n$ avec c une constante incluse dans $]0, 1[$ (par exemple $0,95$) et n la taille du groupe.

Au niveau d'un groupe, toutes les signatures sont d'égale importance, nous répartissons donc de manière équitable les poids attribués au groupe : chacune signature reçoit un poids $p_g = \frac{P}{n} = \frac{c^n}{n}$ qui dépend donc doublement de la taille du groupe.

Seconde étape : calcul global des poids. Une fois l'étape précédente réalisée, nous additionnons les poids obtenus par une même signature σ dans les différents groupes où elle apparaît pour calculer son score $s_\sigma = \Sigma_g p_g$.

Les meilleures signatures, à savoir celles qui se trouvent dans plusieurs groupes et dans des groupes de petite taille, reçoivent alors un score s_σ plus élevé.

4 Synchronisation lexique-grammaire

Cette technique permet à une grammaire d'exprimer ses attentes pour les formes suspectes. Si elle n'est pas parfaite, les représentations qu'elle produit ainsi que les signatures que l'on en extrait ne le sont pas non plus. En fait, dans le cas où la grammaire est parfaite, nous pouvons qualifier les suggestions faites par cette technique comme permettant une correction du lexique. Dans le cas inverse, il s'agit alors d'une technique permettant de diminuer le nombre de conflits entre une grammaire et un lexique, i.e. permettant une meilleure « synchronisation » entre le lexique et la grammaire.

Il est à noter qu'un ensemble de signatures incorrectes représente une source d'informations intéressante sur les manques et incorrections d'une grammaire.

5 Résultats

Le travail présenté ici, tout comme la technique de fouille d'erreurs, est un mécanisme de retour sur erreurs. Ce terme désigne des mécanismes réutilisant les erreurs produites par un programme afin d'améliorer automatiquement ou semi-automatiquement sa qualité. De manière à garantir que l'origine des erreurs produites est effectivement le programme, les données ana-

lysées doivent être fiables. Dans le cas présent, les erreurs sur lesquelles nous travaillons sont issues d'une campagne d'analyse d'un corpus MD de 331 000 phrases extraites du *Monde diplomatique* réalisée durant la validation de la technique de fouille d'erreurs (Sagot & Villemonte de La Clergerie, 2006).

Le lexique que nous cherchons à améliorer est le *Lefff* (***Le**xique des **f**ormes **f**léchies du **f**rançais*) (Sagot *et al.*, 2006). En partie acquis automatiquement, ce lexique morpho-syntaxique à large couverture du français est en constant développement et possède, à l'heure actuelle, plus de 520 000 entrées. La grammaire FRMG (Thomasset & Villemonte de La Clergerie, 2005) que nous utilisons est une grammaire hybride TAG/TIG avec décorations. Elle est construite à partir d'une *méta-grammaire* plus abstraite qui produit un ensemble de 134 arbres très factorisés. Malgré son très faible nombre d'arbres, sa factorisation lui permet de couvrir un grand nombre de cadres de catégorisation pour les verbes, la passivation, les extractions (relatives, interrogatives, clivées), certaines inversions du sujet, certaines constructions à verbe support (« faire attention à »). Néanmoins, nombre de phénomènes ne sont pas encore traités (comme la sous-catégorisation sur les adjectifs et les noms). La grammaire FRMG couplée à *Lefff* assurait en 2005 une couverture de l'ordre de 41% sur le corpus MD.

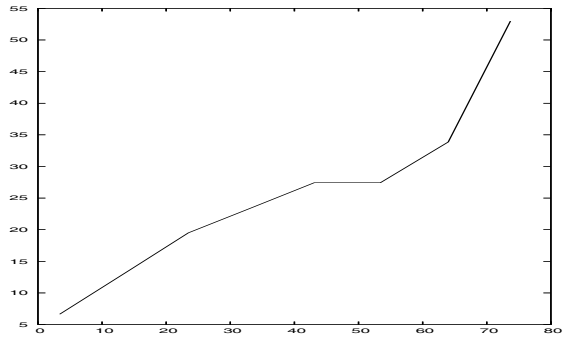


FIG. 2 – Taux de réussite des réanalyses (axe Y) en fonction des taux de suspicion (axe X)

5.1 Exactitude de la détection automatique des formes suspectes

La courbe de la figure 2 nous permet d'observer une corrélation très nette entre les taux de réussite des réanalyses et les taux de suspicion des formes. Cela atteste la validité des informations produites par l'étape précédente de fouille d'erreurs.

Les valeurs présentées par cette courbe sont en réalité des moyennes calculées après un regroupement des formes suspectes par intervalle de taux de suspicion. Sans cela, la courbe présente des variations rendant difficile son observation.
Ces variations s'expliquent principalement par le fait que certaines formes ont été suspectées à la place de la grammaire ce qui explique que leur échange avec des jokers n'ait rien apporté. En effet, certaines formes ont une affinité marquée pour des constructions spécifiques ; par exemple une inversion du sujet en présence de l'adjectif 'rare' comme dans « Rares sont ceux qui tentent

d'en sortir. » ou 'nombreux' dans « Nombreux sont ceux qui refusent. » [2]. Ces formes ont alors payé cette affinité par une suspicion injustement élevée à leur égard.
Une autre raison moins importante expliquant ces variations est que l'utilisation de jokers augmente sensiblement le taux de *timeout* pour les phrases, et cela même en ayant imposé une limite de 40 mots sur la longueur des phrases analysées.

Puisque ces deux phénomènes s'observent à tous les niveaux de taux de suspicion, le regroupement des valeurs par intervalle a permis d'en diminuer l'influence sur la courbe de la figure 2.

Toujours dans une optique de retour sur erreurs, notons qu'il est tentant de voir les phrases des suspects forts avec de faibles taux de réanalyse comme indiquant des manques de la grammaire. Il serait alors intéressant de les analyser au moyen d'un système d'inférence grammaticale.

5.2 Évaluation de la qualité des signatures

Afin d'évaluer la qualité des signatures produites, nous avons ordonné les formes suspectes en accord avec le calcul suivant : $M_f = S_f.ln(NS_f)$, S_f étant le taux de suspicion d'une forme et NS_f le nombre de phrases associées[3]. Nous avons ensuite examiné nombre d'entre elles à travers une interface Web (Fig. 3) nous permettant d'accéder, pour chaque forme, aux jokers testés, aux taux de réanalyses obtenus, aux phrases testées et aux meilleures signatures retenues. De même, elle nous permet de laisser des commentaires et de soumettre des requêtes au lexique et à l'analyseur. À terme, cette interface a vocation à être utilisée par des linguistes.

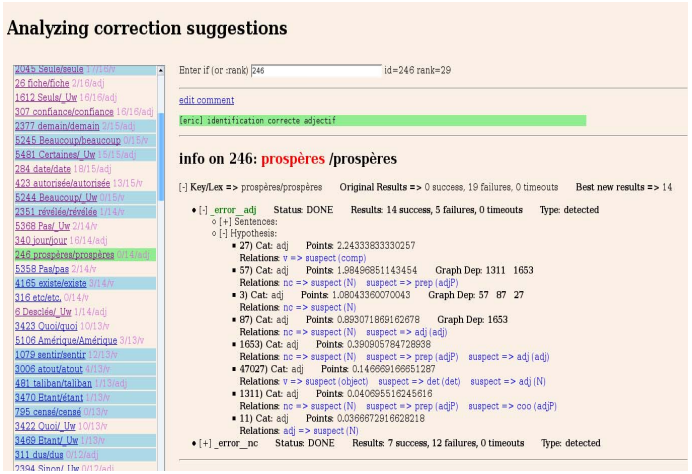


FIG. 3 – Interface d'exploration des signatures

Lors de l'étude des meilleures signatures, certains doutes ont été confirmés : notre technique manque de maturité. Nous avons identifié un certain nombre de phénomènes nous empêchant de correctement quantifier la qualité des signatures. Cependant, nous savons déjà comment faire face à la plupart (voir Sect 6).

[2] Ces exemples reflètent aussi le style recherché du corpus journalistique étudié !
[3] Un fort taux de réanalyse sur un nombre réduit de phrases est peu significatif.

Toutefois, dans bien des cas, nous avons obtenu des résultats pertinents et instructifs qui nous ont permis d'améliorer nos outils (et pas seulement notre lexique). Par exemple, on retrouve la bonne signature comme dans le cas de « prospères » où l'on retrouve un usage d'adjectif épithète (joker + signature), alors qu'il n'existe que comme verbe dans *Lefff*. Pour la forme verbale « révéler » , les hypothèses font ressortir qu'elle attend bien un argument attributif (« ce choix pourrait se révéler catastrophique. ») mais qu'il lui manque le côté réflexif, à cause de constructions prépositionnelles comme « contraint de révéler X » ou « penser à révéler X » .

Bien que devant encore mûrir, notre approche s'est montrée viable. Nous continuerons à la développer afin d'obtenir un outil pleinement fonctionnel. L'achèvement de certaines améliorations donnera notamment lieu à de nouvelles campagnes de calcul.

6 Développements futurs

Durant nos expériences, nous avons pu établir une liste de problèmes à traiter et des solutions pour les résoudre :

- Il est très fréquent de pouvoir appliquer plusieurs productions grammaticales à une suite de formes, surtout si la catégorie syntaxique d'une de ces formes varie (comme pour les jokers). Cependant, ces productions n'ont pas les mêmes fréquences d'utilisations et par conséquent, les signatures qui en résultent ne représentent pas la même quantité d'information utile. De telles données sur les fréquences d'utilisation nous seraient utiles afin de pondérer les signatures et de diminuer l'ingérence de signatures « parasites » dans les résultats.
- Les signatures doivent être nettoyées pour éliminer l'adjonction de certains adjoints (gouvernés par les suspects) qui ne sont pas primordiaux pour caractériser ceux-ci. Cela nous permettrait de consolider des signatures actuellement séparées par des adjoints inutiles. Néanmoins, à ce stade, il n'est pas toujours évident de juger de l'importance d'un adjoint.
- Il nous faut regrouper les formes par famille de lemmes sous-jacents de manière à augmenter la variabilité des contextes testés et ainsi cerner ce qu'ils ont en commun. Néanmoins, il faut garder à l'esprit que certains problèmes ne se manifestent que pour quelques formes, par exemple une mauvaise attribution de l'auxiliaire à utiliser pour des participes passés (exemple de « larvé » faussement listé dans *Lefff* comme utilisant l'auxiliaire « avoir »). Nous avons aussi mentionné que, parfois, le problème résulte du manque dans le lexique d'un des lemmes possibles pour une forme suspecte.
- Il nous faut regrouper les signatures qui traduisent en fait un même phénomène syntaxique sous des aspects différents ; comme par exemple : le sujet et autres arguments verbaux ont diverses réalisations (nominales, cliticisées, pronoms relatifs, pronoms interrogatifs), ou encore un verbe avec objet sous forme active et passive. Le regroupement des formes par lemme est susceptible d'aider.
- Certaines formes suspectes donnent des signatures équivalentes aux informations syntaxiques déjà présentes dans le lexique. Ce genre de cas implique que les signatures sont incomplètes. À l'heure actuelle, elles manquent principalement d'informations morphologiques. L'intégration de ces informations déjà présentes dans les forêts de dépendances, mais non encore exploitées, représente donc la prochaine étape dans l'amélioration du modèle des signatures.
- Certaines formes ont été injustement suspectées à cause de leur affinité avec des constructions syntaxiques non gérées par la grammaire. L'utilisation de plusieurs analyseurs syntaxiques avec des grammaires différentes durant l'étape préalable de fouille d'erreurs permettraient éventuellement de filtrer une partie des formes suspectes.

– Les signatures sont composées d'un ensemble de dépendances syntaxiques entre les mots et le joker dans les représentations générées d'une phrase. Ces signatures dépendent directement de la grammaire utilisée et peuvent être difficiles à comprendre pour une personne non familière avec ce formalisme. Un effort doit donc être réalisé pour les traduire vers une représentation indépendante de la grammaire et plus humainement compréhensible.

7 Conclusion

Les expériences présentées confirment en premier lieu la capacité de la technique de fouille d'erreurs à identifier de bonnes formes suspectes. Leur transformation en jokers augmente le taux de réanalyses réussies de manière coordonnée avec le taux de suspicion d'une forme.

En second lieu, elles valident la faisabilité d'un mécanisme automatique de suggestion de corrections lexicales sur les formes suspectes (i.e. sur les lemmes sous-jacents). Elles montrent qu'il est également possible d'obtenir du retour d'information sur des manques grammaticaux.

Néanmoins, un travail reste encore à faire pour affiner la qualité des corrections suggérées en distinguant mieux l'essentiel de l'accessoire dans les signatures, notamment à travers des améliorations introduites précédemment.

Références

ABEILLÉ A. (1993). *Les nouvelles syntaxes, grammaire d'unification et analyse du français*. Armand Colin.

BRENT M. R. (1993). From grammar to lexicon : unsupervised learning of lexical syntax. *Computational Linguistic*, **19**(2), 243–262.

CRYSMANN B., FRANK A., KIEFER B., KRIEGER H.-U., MÜLLER S., NEUMANN G., PISKORSKI J., SCHÄFER U., SIEGEL M., USZKOREIT H. & XU F. (2002). An integrated architecture for shallow and deep processing. In *Proceedings of the 40th Annual Meeting of the ACL*, p. 441–448.

GROVER C. & LASCARIDES A. (2001). XML-based data preparation for robust deep parsing. In *Meeting of the Association for Computational Linguistics*, p. 252–259.

HORIGUCHI K., TORISAWA K. & TSUJII J. (1995). Automatic acquisition of content words using an HPSG-based parser. In *Proceedings of NLPRS'95.*

SAGOT B. & BOULLIER P. (2005). From raw corpus to word lattices : robust pre-parsing processing. In *Proceedings of L&TC*, Poznan, Pologne.

SAGOT B., CLÉMENT L., VILLEMONTE DE LA CLERGERIE É. & BOULLIER P. (2006). The Lefff 2 syntactic lexicon for french : architecture, acquisition, use. In *Proceedings of LREC'06*.

SAGOT B. & VILLEMONTE DE LA CLERGERIE É. (2006). Trouver le coupable : Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Proceedings of TALN'06*, p. 287–296.

SCHMID H. (1999). Probabilistic part-of-speech tagging using decision trees. *IMS-CL.*

THOMASSET F. & VILLEMONTE DE LA CLERGERIE É. (2005). Comment obtenir plus des méta-grammaires. In *Proceedings of TALN'05*, Dourdan, France : ATALA.

VAN NOORD G. (2004). Error mining for wide-coverage grammar engineering. In *Proceedings of ACL 2004*, Barcelone, Espagne.

Ambiguïté de portée et approche fonctionnelle des grammaires d'arbres adjoints

Sylvain POGODALLA
LORIA/INRIA Lorraine
sylvain.pogodalla@loria.fr

Résumé. En s'appuyant sur la notion d'arbre de dérivation des Grammaires d'Arbres Adjoints (TAG), cet article propose deux objectifs : d'une part rendre l'interface entre syntaxe et sémantique indépendante du langage de représentation sémantique utilisé, et d'autre part offrir un noyau qui permette le traitement sémantique des ambiguïtés de portée de quantificateurs sans utiliser de langage de représentation sous-spécifiée.

Abstract. Relying on the derivation tree of the Tree Adjoining Grammars (TAG), this paper has to goals : on the one hand, to make the syntax/semantics interface independant from the semantic representation language, and on the other hand to propose an architecture that enables the modeling of scope ambguities without using underspecified representation formalisms.

Mots-clés : interface syntaxe et sémantique, sémantique formelle, grammaires d'arbres adjoints, grammaires catégorielles.

Keywords: syntax/semantics interface, formal semantics, tree adjoining grammars, categorial grammars.

1 Introduction

La notion d'arbre de dérivation dans les grammaires d'arbres adjoints (TAG) (Joshi & Schabes, 1997; Abeillé, 1993) est censée représenter les dépendances entre les différents items lexicaux d'une phrase. À ce titre, l'arbre de dérivation apparaît comme le candidat privilégié pour réaliser le transfert structurel entre la syntaxe et la sémantique de manière compositionnelle. Or, sa représentation ne rendant pas explicite certains liens, il a été proposé, afin de le rendre opérationnel dans le cadre du calcul de la représentation sémantique, soit de l'étendre (Kallmeyer, 2002; Joshi *et al.*, 2003), soit de ne pas l'utiliser et de calculer la représentation sémantique directement sur l'arbre dérivé (Frank & van Genabith, 2001; Gardent & Kallmeyer, 2003; Gardent, 2007).

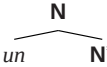
Cet article propose d'utiliser la notion d'arbre de dérivation telle qu'introduite dans (Pogodalla, 2004). En effet, cette notion, qui précise simplement la notion originale, y est montrée comme adéquate pour la représentation des dépendances longue distance. Néanmoins, le langage de représentation sémantique qui est utilisé est un formalisme sous-spécifié. Ces derniers posent parfois problème, comme dans le cas de la coordination de groupes nominaux quantifiés (Willis, 2007). De plus, nous voulons un cadre général qui laisse à l'utilisateur le choix d'utiliser ou

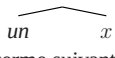
non de tels formalismes, tout en gardant la possibilité de modéliser les ambiguïtés. Ainsi, nous utilisons un formalisme plus proche de celui proposé par Montague (Montague, 1974) et une architecture qui permet de traiter des phénomènes d'ambiguïté. Nous nous appuyons sur les Grammaires Catégorielles Abstraites (ACG) (de Groote, 2001), et, *tout en gardant un seul arbre dérivé*, nous montrons comment le principe d'élévation de type des grammaires catégorielles permet d'obtenir plusieurs lectures sémantiques.

Dans les deux prochaines sections, nous présentons l'arbre de dérivation de (Pogodalla, 2004) sur des exemples. Puis nous définissons dans la section 4 la notion d'ACG et les architectures qu'elle rend possible pour l'interface entre la syntaxe et la sémantique. La section 5 met finalement en œuvre une telle architecture pour modéliser l'ambiguïté de portée des quantificateurs.

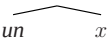
2 Lecture fonctionnelle de l'arbre dérivé

La présentation habituelle des arbres élémentaires des TAG en donne une vision relativement statique, les opérations de substitution et d'adjonction se chargeant de donner la dynamique du langage en combinant selon certaines règles les arbres entre eux. Dans cette section, nous nous proposons d'intégrer cette notion de dynamique aux arbres élémentaires eux-mêmes, en décrivant comment chacun prend part aux opérations de substitution et d'adjonction. Cette description se fait sur base d'exemples.

Soit l'arbre auxiliaire suivant : 

. Lorsqu'il est adjoint à un autre arbre à un nœud $\mathbf{N}_0$, cet arbre remplace son propre nœud $\mathbf{N}^*$ par le sous-arbre de racine $\mathbf{N}_0$. Si l'on appelle x ce sous-arbre, on peut donc considérer l'arbre auxiliaire comme une fonction qui transforme un arbre x en un nouvel arbre 

. Soit, avec la notation à l'aide du λ-calcul, une représentation de cet arbre par le terme suivant :

$$c_{un} = \lambda x.$$ 

Considérons maintenant l'arbre initial suivant : 

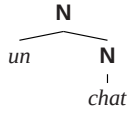
. Cet arbre peut se voir adjoindre un arbre auxiliaire au nœud **N**. Dans ce cas, il donnera comme argument à cet arbre auxiliaire (on a vu que l'arbre auxiliaire peut être décrit comme étant une fonction qui prend un arbre en argument et retourne un arbre) le sous arbre 

(dans le cas présent, le sous-arbre paramètre est l'arbre tout entier car l'adjonction a lieu au nœud racine).

On peut donc représenter l'arbre initial comme une fonction qui prend comme paramètre un arbre auxiliaire, c'est-à-dire *une fonction des arbres dans les arbres*. Soit, avec la notation en λ-calcul :

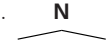
$$\lambda a.a($$ 

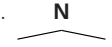
$$)$$

On constate alors que l'opération d'adjonction qui permet d'obtenir l'arbre 

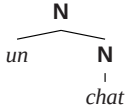
est décrite par l'application de la fonction (du terme) c_{chat} au terme c_{un}. En effet :

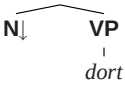
$c_{chat}\,c_{un} = (\lambda a.a(\,$ 

$\,))(\lambda x.\,$ 

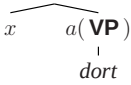
$\,) \rightarrow_\beta (\lambda x.\,$ 

$\,)(\,$ 

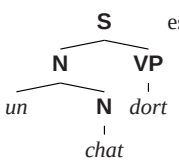
$\,) \rightarrow_\beta$ 

On peut finalement avoir un arbre qui combine la possibilité de subir une adjonction et une substitution. Prenons par exemple l'arbre initial suivant : 

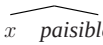
. Sa particularité est qu'il attend un arbre qui peut être substitué au nœud **N** d'une part, et qu'il peut subir une adjonction au nœud **VP**. On choisit donc de le représenter comme une fonction qui prend en premier argument un arbre auxiliaire, c'est-à-dire une fonction, et en deuxième argument un arbre x qui est celui qui est substitué au nœud **N**. On peut alors le représenter de la manière suivante :

$c_{dort} = \lambda a x.$ 

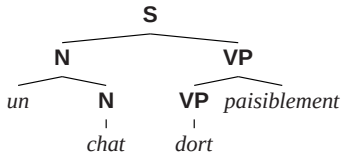
Bien entendu, il est possible qu'aucune adjonction n'ait lieu sur le nœud **VP**[1]. Dans l'optique que nous avons choisie, cela signifie que la fonction qui a été adjointe est l'identité $I = \lambda x.x$.

L'arbre dérivé 

est alors représenté par le terme $c_{dort}\ I\ (c_{chat}\,c_{un})$.

Avec une représentation adéquate de l'adverbe, par exemple $c_{paisiblement} = \lambda x.$ 

, on peut également construire l'arbre dérivé représenté par le terme $c_{dort}\,c_{paisiblement}\,(c_{chat}\,c_{un})$:



[1] Pour des raisons de clarté dans la présentation, nous avons omis la possibilité d'une adjonction sur le nœud **S**, et donc supprimé le paramètre correspondant. On voit également par là comment interdire des adjonctions.

Si l'on appelle γ le type des arbres, on voit que l'on a les constantes et le typage suivants :

$$\begin{array}{ll} c_{un} & : \gamma \multimap \gamma \\ c_{chat} & : (\gamma \multimap \gamma) \multimap \gamma \\ c_{dort} & : (\gamma \multimap \gamma) \multimap \gamma \multimap \gamma \end{array} \qquad \begin{array}{ll} c_{paisiblement} & : \gamma \multimap \gamma \\ I & : \gamma \multimap \gamma \end{array}$$

où $\multimap$ désigne l'implication linéaire[2].

3 Rôle de l'arbre de dérivation

En typant les constantes représentant les arbres auxiliaires et initiaux de cette manière, nous perdons toutefois une information importante : les arbres ont tous le même type γ, et aucune distinction n'est faire entre eux. Ainsi, la composition $c_{chat}c_{paisiblement}$ serait tout à fait licite. C'est pourquoi nous allons donner aux constantes un type plus abstrait[3], correspondant aux non-terminaux qui étiquettent leur racine. Nous nous donnons donc les types de base suivants : **VP**, **S**, **V**, **N** ainsi que les types qui correspondent aux racines des nœuds auxiliaires : $\mathbf{VP}_\mathsf{A}$, $\mathbf{S}_\mathsf{A}$, $\mathbf{V}_\mathsf{A}$, $\mathbf{N}_\mathsf{A}$.

Ainsi, en reprenant les exemples ci-dessus et en introduisant de nouvelles constantes, nous avons les typages suivants :

$$\begin{array}{ll} C_{dort} & : \mathbf{VP}_\mathsf{A} \multimap \mathbf{N} \multimap \mathbf{S} \\ C_{chat} & : \mathbf{N}_\mathsf{A} \multimap \mathbf{N} \end{array} \qquad \begin{array}{ll} C_{un} & : \mathbf{N}_\mathsf{A} \\ C_{paisiblement} & : \mathbf{VP}_\mathsf{A} \\ I_{\mathbf{VP}} & : \mathbf{VP}_\mathsf{A} \end{array}$$

On peut alors avoir le terme $C_{dort}I_{\mathbf{VP}}(C_{chat}C_{un})$, de type **S**, tandis que le terme $C_{chat}C_{paisiblement}$ n'est pas typable. Il reste à établir le lien avec le terme $c_{dort}(c_{chat}c_{un})$ de la section précédente. Cela se fait par une fonction de conversion $:=_{\text{syntax}}$, le *lexique*, qui convertit les types et les constantes ainsi :

S	$:=_{\text{syntax}}$	γ	C_{dort}	$:=_{\text{syntax}}$	c_{dort}
VP	$:=_{\text{syntax}}$	γ	C_{un}	$:=_{\text{syntax}}$	c_{dort}
N	$:=_{\text{syntax}}$	γ	C_{chat}	$:=_{\text{syntax}}$	c_{chat}
$\mathbf{N}_\mathsf{A}$	$:=_{\text{syntax}}$	$\gamma \multimap \gamma$	$C_{paisiblement}$	$:=_{\text{syntax}}$	$c_{paisiblement}$
$\mathbf{VP}_\mathsf{A}$	$:=_{\text{syntax}}$	$\gamma \multimap \gamma$	I_X	$:=_{\text{syntax}}$	$\lambda x.x$ pour tout type X

TAB. 1 – Définition du lexique

On alors :

$$C_{dort}I_{\mathbf{VP}}(C_{chat}C_{un}) \quad :=_{\text{syntax}} \quad c_{dort}I(c_{chat}c_{un})$$

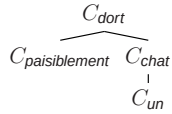
et

$$C_{dort}C_{paisiblement}(C_{chat}C_{un}) \quad :=_{\text{syntax}} \quad c_{dort}c_{paisiblement}(c_{chat}c_{un})$$

Si l'on adopte une représentation arborescente des λ-termes (il n'y a pas d'abstraction), on peut représenter $C_{dort}C_{paisiblement}(C_{chat}C_{un})$ par :

[2]Nous ne dirons rien du calcul logique sous-jacent, ni de la manière d'introduire de la non-linéarité avec l'implication intuitionniste habituelle $\rightarrow$. Nous renvoyons les lecteurs intéressés à (de Groote, 2001; Pogodalla, 2004).

[3]Car pouvant être réalisé, ou instancié, de différentes manières : arbre (γ) pour la syntaxe, mais aussi individu (e), prédicat ($e \multimap t$), etc. pour la sémantique.



Cet arbre rappelle très précisément *l'arbre de dérivation* tel qu'il est défini classiquement dans les TAG. En fait, il s'agit de la même notion où sont cependant explicités :
– l'ordre des arguments, qui doit être le même pour la constante qui est représentée dans l'arbre de dérivation et pour la constante qui lui est associée dans les arbres dérivés. Le choix est libre, mais une fois qu'il est fait, il doit être cohérent ;
– l'ordre des adjonctions lors d'une dérivation. Contrairement à la notion classique, où cet ordre n'est pas précisé, le résultat étant le même, ici l'ordre des opérations est spécifié. Cela ne change pas le pouvoir expressif, cela permet par contre de doter les TAG d'une sémantique compositionnelle basée sur l'arbre de dérivation.

Cette manière de représenter les arbres dérivés, les arbres de dérivation, et les relations qu'il y a entre eux, correspond en fait à la modélisation des TAG dans le formalisme des ACG.

4 Modélisation des TAG dans les ACG

Nous ne reprenons pas ici le détail la modélisation systématique des TAG dans les ACG, donné dans (de Groote, 2002; Pogodalla, 2004). Nous allons simplement donner les définitions précises des ACG qui ont été mises en œuvre dans les exemples précédents, afin d'en tirer l'architecture générale que nous utiliserons pour modéliser les ambiguïtés de portée des quantificateurs.

Une ACG définit deux langages : un *langage abstrait*, qui peut être vu comme un ensemble abstrait de structures grammaticales, et un *langage objet*, représentant les formes réalisées des structures abstraites, qu'elle met en relation. Ici, le langage abstrait correspond à la structure grammaticale que l'on veut manipuler : l'arbre de dérivation. Dans l'exemple précédent, il est mis en relation avec le langage objet des arbres dérivés grâce au *lexique*.

Definition 1 (Signature d'ordre supérieur). *Une* signature d'ordre supérieure *est un triplet* $\Sigma = \langle A, C, \tau \rangle$ *où :*
– A *est un ensemble de types atomiques ;*
– C *est un ensemble fini de constantes ;*
– $\tau : C \rightarrow \mathcal{T}(A)$ *qui assigne à chaque constante de* C *un type de* $\mathcal{T}(A)$ *où* $\mathcal{T}(A) ::= A | \mathcal{T}(A) \multimap \mathcal{T}(A)$.
On appel Λ_Σ *l'ensemble des* λ*-termes que l'on peut construire avec la signature* Σ.

Ainsi, dans l'exemple précédent, nous avions deux signatures d'ordre supérieur. La première contenait les types atomiques $\mathbf{S}$, $\mathbf{N}$, $\mathbf{VP}_{\mathsf{A}}$... et les constantes C_{chat}, C_{un}... tandis que la deuxième signature d'ordre supérieur contenait l'unique type atomique γ et les constante c_{chat}, c_{un}...

Definition 2 (Lexique). *Étant données une signature d'ordre supérieur* $\Sigma_1 = \langle A_1, C_1, \tau_1 \rangle$ *et une signature d'ordre supérieur* $\Sigma_2 = \langle A_2, C_2, \tau_2 \rangle$*, un* lexique $:=$ de Σ_1 vers Σ_2 *est défini par la donnée de* $\overset{\tau}{:=}$ *et* $\overset{c}{:=}$ *tels que :*

– $:\overset{\tau}{=}\ : A_1 \rightarrow \mathcal{T}(A_2)$ *est une fonction d'interprétation des types atomiques de* Σ_1 *comme des types implicatifs construits à partir de* A_2*. On appellera* $:\overset{\tau}{=}$ *également son extension homomorphique à tous les types de* $\mathcal{T}(A_1)$ *;*
– $:\overset{c}{=}\ : C_1 \rightarrow \Lambda_{\Sigma_2}$ *est une fonction d'interprétation des constantes de* Σ_1 *comme des* λ*-termes construits à partir de* Σ_2*. On appellera* $:\overset{c}{=}$ *également son extension homomorphique à tous les termes de* Λ_{Σ_1} *;*
– *les fonctions d'interprétation sont compatibles avec la relation de typage, c'est-à-dire que pour tout* $c \in C_1$ *et* $t : \alpha \in \Lambda_{\Sigma_2}$ *tels que* $c :\overset{c}{=} t$*, alors* $\tau_1(c) :\overset{\tau}{=} \alpha$ *(le type de l'image de* c *est l'image du type de* c*).*

Dans la suite, on utilisera sans ambiguïté $:=$ *pour* $:\overset{\tau}{=}$ *ou* $:\overset{c}{=}$*.*

Le tableau 1 définit bien un lexique. La colonne de gauche donne l'interprétation des types atomiques (on remarquera avec l'interprétation du type $\mathbf{VP}_{\mathsf{A}}$ que l'interprétation d'un type atomique peut être un type non atomique). La colonne de droite donne l'interprétation des constantes.

Definition 3 (Grammaire catégorielle abstraite). *Une* grammaire catégorielle abstraite *est un quadruplet* $\mathcal{G} = \langle \Sigma_1, \Sigma_2, :=, s \rangle$ *où :*
– Σ_1 *est une signature d'ordre supérieure, et* Σ_2 *une signature d'ordre supérieure. Ils sont appelés* vocabulaire abstrait *et* vocabulaire objet *;*
– $:=\ : \Sigma_1 \rightarrow \Sigma_2$ *est un lexique ;*
– s *est un type atomique du vocabulaire abstrait, appelé le* type distingué *de la grammaire.*

Definition 4 (Langages abstrait et objet). *Soit* $\mathcal{G} = \langle \Sigma_1, \Sigma_2, :=, s \rangle$ *une grammaire catégorielle abstraite.*

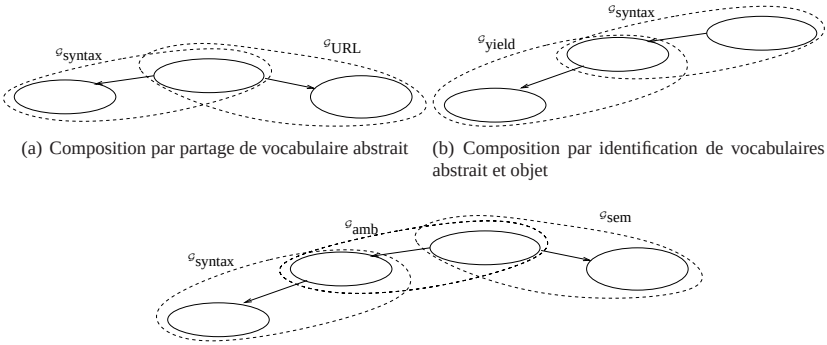
1. *Le* langage abstrait $\mathcal{A}(\mathcal{G})$ *engendré par* $\mathcal{G}$ *est défini par* $\mathcal{A}(\mathcal{G}) = \{t \in \Lambda_{\Sigma_1} | t : s\}$
2. *Le* langage objet $\mathcal{O}(\mathcal{G})$ *engendré par* $\mathcal{G}$ *est défini par* $\mathcal{O}(\mathcal{G}) = \{t \in \Lambda_{\Sigma_2} | \exists u \in \mathcal{A}(\mathcal{G}) \text{ avec } u := t\}$

Ainsi, les termes pris en exemple appartiennent bien aux vocabulaires abstrait et objet. Il est à noter que cette définition permet d'éviter que le terme $c_{\mathit{dort}}(c_{\mathit{chat}} c_{\mathit{paisiblement}})$, qui est bien un arbre (de type γ), appartienne effectivement au langage objet des arbres dérivés. En effet, il serait l'image de $C_{\mathit{dort}}(C_{\mathit{chat}} C_{\mathit{paisiblement}})$ qui n'est pas de type $\mathbf{S}$ (ce terme n'est même pas typable) et qui n'appartient donc pas au langage abstrait des arbres de dérivation.

La définition des ACG permet de considérer différents types d'architecture. Par exemple, si deux ACG partagent le même vocabulaire abstrait, on aura le schéma de composition de la figure 1(a). C'est par exemple celui adopté dans (Pogodalla, 2004) pour doter les TAG d'une représentation sémantique sous-spécifiée.

On peut également composer deux ACG en faisant que le vocabulaire objet de l'une soit également le vocabulaire abstrait de l'autre (figure 1(b)). C'est par exemple le cas si l'on veut considérer le lien entre les arbres dérivés, cette fois vus comme un langage abstrait, et leur production (*yield* en anglais) comme langage de chaîne.

Bien entendu, on peut mélanger ces deux types de composition. La modélisation que nous proposons pour les phénomènes d'ambiguïté de portée des quantificateurs repose sur le schéma de la figure 1(c). Dans tous les cas, on retrouve un schéma classique du TAL, même si la relation est décrite par un autre formalisme : celui de la composition de *transducer*.



(a) Composition par partage de vocabulaire abstrait

(b) Composition par identification de vocabulaires abstrait et objet

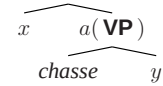
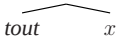
(c) Composition par combinaison des deux modes

FIG. 1 – Exemples d'architectures possibles

5 Composition d'ACG et modélisation des ambiguïtés de portée

Proposition. Notre objectif est de proposer pour les TAG un cadre dans lequel modéliser les ambiguïtés de portée sans utiliser de formalisme sous-spécifié (contrairement à (Pogodalla, 2004)), tout en gardant la contrainte d'avoir un unique arbre dérivé auquel peuvent être associées plusieurs représentations sémantiques. Pour l'architecture que nous proposons, il nous faut définir deux nouvelles ACG. La première, $\mathcal{G}_{\text{amb}}$, permettra d'associer à un arbre de dérivation unique deux structures plus profondes. La seconde, $\mathcal{G}_{\text{sem}}$, correspondra à la réalisation dans un langage de formes logiques du type de Montague de ces structures plus profondes.

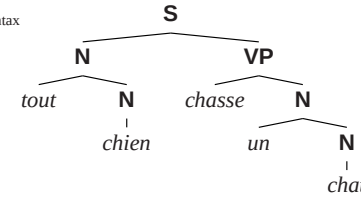
On pourra penser que ces différents niveaux ajoutent de la complexité conceptuelle. Nous pensons pour notre part que cela permet de modulariser les difficultés (en ne traitant qu'à leurs niveaux respectifs la syntaxe, avec $\mathcal{G}_{\text{syntax}}$, et la sémantique, avec $\mathcal{G}_{\text{amb}}$ et $\mathcal{G}_{\text{sem}}$). Par ailleurs, nous avons vu que cette notion de composition est déjà présente en TAL avec l'utilisation de *transducer* et de leur composition, pour le traitement de la morphologie par exemple. Si au final seule la relation entre les langages d'entrée et sortie nous intéresse, il est tout à fait possible de compiler deux ACG, par exemple $\mathcal{G}_{\text{amb}}$ et $\mathcal{G}_{\text{syntax}}$ en une seule, leur *composée*.

Supposons que la grammaire $\mathcal{G}_{\text{syntax}}$ contiennent également les arbres suivants : $c_{\textit{chasse}} = \lambda axy.$  , $c_{\textit{chien}} = \lambda a.a($  $)$ et $c_{\textit{tout}} = \lambda x.$  , ainsi que les

constantes $C_{\textit{chien}} : \mathbf{N}_{\mathsf{A}} \multimap \mathbf{N}$, $C_{\textit{chasse}} : \mathbf{VP}_{\mathsf{A}} \multimap \mathbf{N} \multimap \mathbf{N} \multimap \mathbf{S}$ et $C_{\textit{tout}} : \mathbf{N}_{\mathsf{A}}$ mis en relation par le lexique de la manière suivante : $C_{\textit{chasse}} :=_{\text{syntax}} c_{\textit{chasse}}$, $C_{\textit{chien}} :=_{\text{syntax}} c_{\textit{chien}}$ et $C_{\textit{tout}} :=_{\text{syntax}} c_{\textit{tout}}$.

Alors l'arbre de dérivation correspondant à l'arbre dérivé

$c_{chasse}(c_{chien}c_{tout})(c_{chat}c_{un}) :=_{\text{syntax}}$



est $t_0 = C_{chasse} I_{\mathbf{VP}} (C_{chien} C_{tout})(C_{chat} C_{un})$.

Définissons maintenant $\mathcal{G}_{\text{amb}}$, dont le vocabulaire objet est le vocabulaire abstrait de $\mathcal{G}_{\text{syntax}}$, et dont le vocabulaire abstrait contient les mêmes symboles de type que le vocabulaire objet mais les constantes typées $D_{chasse} : \mathbf{VP}_{\mathsf{A}} \multimap \mathbf{N} \multimap \mathbf{N} \multimap \mathbf{S}$, $D_{chien} : \mathbf{N}_{\mathsf{A}} \multimap (\mathbf{N} \multimap \mathbf{S}) \multimap \mathbf{S}$, $D_{chat} : \mathbf{N}_{\mathsf{A}} \multimap (\mathbf{N} \multimap \mathbf{S}) \multimap \mathbf{S}$, $D_{tout} : \mathbf{N}_{\mathsf{A}}$, $D_{un} : \mathbf{N}_{\mathsf{A}}$ et $I^D_{\mathbf{VP}} : \mathbf{VP}_{\mathsf{A}}$. Le lexique $:=_{\text{amb}}$ est tel que pour tout type X, $X :=_{\text{amb}} X$ et :

$$\begin{array}{llll} D_{chasse} & :=_{\text{amb}} & C_{chasse} & \qquad D_{chien} \quad :=_{\text{amb}} \quad \lambda aP.P(C_{chien}\, a) \\ D_{tout} & :=_{\text{amb}} & C_{tout} & \qquad D_{chat} \quad :=_{\text{amb}} \quad \lambda aP.P(C_{chat}\, a) \\ D_{un} & :=_{\text{amb}} & C_{un} & \qquad I^D_{\mathbf{VP}} \quad :=_{\text{amb}} \quad I_{\mathbf{VP}} \end{array}$$

Soit alors les termes :

$$\begin{array}{ll} t_1 & = (D_{chien} D_{tout})(\lambda x.(D_{chat} D_{un})(\lambda y. D_{chasse} I^D_{\mathbf{VP}}\, x\, y)) \\ t_2 & = (D_{chat} D_{un})(\lambda y.(D_{chien} D_{tout})(\lambda x. D_{chasse} I^D_{\mathbf{VP}}\, x\, y)) \end{array}$$

On pourra vérifier que t_1 et t_2 sont bien typés et que $t_1 :=_{\text{amb}} t_0$ et $t_2 :=_{\text{amb}} t_0$. Ainsi, nous avons désormais deux structures profondes (t_1 et t_2) reliées à un seul arbre de dérivation (t_0).

Il nous reste à transformer ces structures en formules logiques à l'aide d'une nouvelle ACG $\mathcal{G}_{\text{sem}}$. Celle-ci partage son vocabulaire abstrait avec $\mathcal{G}_{\text{amb}}$, et, au niveau objet, met en œuvre les types habituels e et t pour les représentations à la Montague. Avec le lexique $:=_{\text{sem}}$[4] suivant[5] :

$$\begin{array}{lll} \mathbf{S} & :=_{\text{sem}} t & \qquad D_{chasse} \quad :=_{\text{sem}} \quad \lambda aso.(a(\lambda x.\mathbf{chasse}\, x\, o))s \\ \mathbf{N} & :=_{\text{sem}} e & \qquad D_{tout} \quad :=_{\text{sem}} \quad \lambda PQ.\forall x.P\, x \Rightarrow Q\, x \\ \mathbf{N}_{\mathsf{A}} & :=_{\text{sem}} (e \to t) \to (e \to t) \to t & \qquad D_{un} \quad :=_{\text{sem}} \quad \lambda PQ.\exists x.P\, x \wedge Q\, x \\ \mathbf{VP}_{\mathsf{A}} & :=_{\text{sem}} (e \to t) \to (e \to t) & \qquad D_{chat} \quad :=_{\text{sem}} \quad \lambda q.q(\lambda x.\mathbf{chat}\, x) \\ & & \qquad D_{chien} \quad :=_{\text{sem}} \quad \lambda q.q(\lambda x.\mathbf{chien}\, x) \\ & & \qquad I^D_{\mathbf{VP}} \quad :=_{\text{sem}} \quad \lambda x.x \end{array}$$

Nous laissons le lecteur vérifier que l'on obtient bien alors les deux lectures :

$$\begin{array}{ll} t_1 & :=_{\text{sem}} \quad \forall x.\mathbf{chien}\, x \Rightarrow (\exists y.\mathbf{chat}\, y \wedge \mathbf{chasse}\, x\, y) \\ t_2 & :=_{\text{sem}} \quad \exists y.\mathbf{chat}\, y \wedge (\forall x.\mathbf{chien}\, x \Rightarrow \mathbf{chasse}\, x\, y) \end{array}$$

Faute de place, nous ne pouvons illustrer également la coordination de groupes nominaux quantifiés avec les constantes $C_{et} : \mathbf{N} \multimap \mathbf{N} \multimap \mathbf{N}$, $D_{et} : ((\mathbf{N} \multimap \mathbf{S}) \multimap \mathbf{S}) \multimap ((\mathbf{N} \multimap$

[4] On suppose présentes dans la signature objet les constantes **chasse** : $e \multimap e \multimap t$, **chien** : $e \multimap t$, **chat** : $e \multimap t$, $\forall : (e \to t) \multimap t$ et $\exists : (e \to t) \multimap t$.

[5] Notons que c'est la présence du paramètre a dans la formule sémantique qui réalise D_{chasse} qui permet, en intégrant la contribution des éventuels sous-arbres adjoints, la prise en compte des dépendances longue distance.

$\mathbf{S}) \multimap \mathbf{S}) \multimap ((\mathbf{N} \multimap \mathbf{S}) \to \mathbf{S})$ et leur réalisation $D_{\mathit{et}} :=_{\text{amb}} \lambda PQr.P(\lambda x.Q(\lambda y.r(C_{\mathit{et}}\, x\, y)))$ et $D_{\mathit{et}} :=_{\text{sem}} \lambda PQr.P\, r \wedge Q\, r$. On aurait par exemple les deux termes

$$\begin{aligned} t_3 &= D_{\mathit{et}}(D_{\mathit{chat}}\, D_{\mathit{tout}})(D_{\mathit{chien}}\, D_{\mathit{un}})(\lambda x.(D_{\mathit{souris}}\, D_{\mathit{une}})(\lambda y.D_{\mathit{chasse}}\, I^{D}_{\mathbf{VP}}\, x\, y)) \\ t_4 &= (D_{\mathit{souris}}\, D_{\mathit{une}})(\lambda y.D_{\mathit{et}}(D_{\mathit{chat}}\, D_{\mathit{tout}})(D_{\mathit{chien}}\, D_{\mathit{un}})(\lambda x.D_{\mathit{chasse}}\, I^{D}_{\mathbf{VP}}\, x\, y)) \end{aligned}$$

qui donneraient les deux lectures attendues pour *tout chat et un chien chassent une souris*. Contrairement au problème soulevé par les représentations sous-spécifiées dans (Willis, 2007), on n'a pas la lecture où *tout chat* a une portée différente de *un chien* vis à vis de la portée de *une souris*. On obtient ainsi une architecture dans laquelle modéliser les phénomènes d'ambiguïté de portée sans imposer l'utilisation de formalismes sous-spécifiés.

Limitations. Actuellement, nous ne savons pas exprimer les contraintes de portée des quantificateurs, telles celles des îlots de portée. Ce problème est comparable à celui rencontré par les grammaires de types logiques. En effet, l'approche proposée ici repose sur le principe de l'élévation de type, qui est à la base de la prise en compte des ambiguïtés de portée dans ces grammaires. Ici, nous avons gardé la contrainte supplémentaire que, bien entendu, l'arbre dérivé et l'arbre de dérivation restent uniques. La solution que nous envisageons repose sur une extension du système de type des ACG, et va bien au-delà du sujet de cet article[6].

6 Conclusion

Nous avons montré comment, en se basant sur la définition précise de l'arbre de dérivation de (Pogodalla, 2004), nous pouvons définir un calcul des représentations sémantiques pour les TAG qui ne nécessite pas l'usage de formalismes sous-spécifiés tout en permettant le traitement de l'ambiguïté. Cela nous permet d'une part de renforcer l'indépendance entre le formalisme syntaxique des TAG et le formalisme choisi par l'utilisateur pour la représentation sémantique, et d'autre par de confirmer l'importance de cette notion d'arbre de dérivation. Par ailleurs, notre approche a de forts liens avec les approches de Glue Semantics (Dalrymple, 2001), et la proposition (Frank & van Genabith, 2001) (utilisant les principes de Glue Semantics depuis l'arbre dérivé) pourrait sans doute être reconsidérée avec cette notion d'arbre de dérivation.

Références

ABEILLÉ A. (1993). *Les nouvelles syntaxes*. Paris : Armand Colin Éditeur.

DALRYMPLE M. (2001). *Lexical Functional Grammar*, volume 42 of *Syntax and Semantics series*. Academic Press.

DE GROOTE P. (2001). Towards abstract categorial grammars. In *Association for Computational Linguistics, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference*, p. 148–155.

DE GROOTE P. (2002). Tree-adjoining grammars as abstract categorial grammars. In *TAG+6, Proceedings of the sixth International Workshop on Tree Adjoining Grammars and Related Frameworks*, p. 145–150 : Università di Venezia.

[6] Faute de place, nous ne pouvons pas exposer comment l'architecture proposée ici permet également de dépasser la limitation mentionnée dans (Pogodalla, 2004) pour les verbes à contrôle.

FRANK A. & VAN GENABITH J. (2001). Glue tag : Linear logic based semantics construction for LTAG - and what it teaches us about the relation between LFG and LTAG. In M. BUTT & T. H. KING, Eds., *Proceedings of the LFG '01 Conference*, Online Proceedings : CSLI Publications. `http://cslipublications.stanford.edu/LFG/6/lfg01.html`.

GARDENT C. (2007). Tree adjoining grammar, semantic calculi and labelling invariants. In (Getzen *et al.*, 2007).

GARDENT C. & KALLMEYER L. (2003). Semantic construction in feature-based tag. In *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*.

J. GETZEN, E. THIJSSE, H. BUNT & A. SCHIFFRIN, Eds. (2007). *Proceedings of the Seventh International Workshop on Computational Semantics, IWCS-7*. Tilburg University.

JOSHI A. K., KALLMEYER L. & ROMERO M. (2003). Flexible composition in ltag : Quantifier scope and inverse linking. In H. BUNT, I. VAN DER SLUIS & R. MORANTE, Eds., *Proceedings of the Fifth International Workshop on Computational Semantics IWCS-5*.

JOSHI A. K. & SCHABES Y. (1997). Tree-adjoining grammars. In G. ROZENBERG & A. SALOMAA, Eds., *Handbook of formal languages*, chapter 2. Springer.

KALLMEYER L. (2002). Using an enriched tag derivation structure as basis for semantics. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+6)*.

MONTAGUE R. (1974). The proper treatment of quantification in ordinary english. In P. PORTNER & B. H. PARTEE, Eds., *Formal Semantics : The Essential Readings*, chapter 1. Blackwell Publishers. 2002 edition.

POGODALLA S. (2004). Computing semantic representation : Towards ACG abstract terms as derivation trees. In *Proceedings of the Seventh International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG+7)*, p. 64–71.

WILLIS A. (2007). NP coordination in underspecified scope representations. In (Getzen *et al.*, 2007).

Évaluer SYNLEX

Ingrid FALK[1], Gil FRANCOPOULO[2], Claire GARDENT[3]
[1] CNRS/ATILF, Nancy
[2] INRIA/LORIA, Nancy
[3] CNRS/LORIA, Nancy
{Ingrid.Falk,Claire.Gardent}@loria.fr,
Gil.Francopoulo@wanadoo.fr

Résumé. SYNLEX est un lexique syntaxique extrait semi-automatiquement des tables du LADL. Comme les autres lexiques syntaxiques du français disponibles et utilisables pour le TAL (LEFFF, DICOVALENCE), il est incomplet et n'a pas fait l'objet d'une évaluation permettant de déterminer son rappel et sa précision par rapport à un lexique de référence. Nous présentons une approche qui permet de combler au moins partiellement ces lacunes. L'approche s'appuie sur les méthodes mises au point en acquisition automatique de lexique. Un lexique syntaxique distinct de SYNLEX est acquis à partir d'un corpus de 82 millions de mots puis utilisé pour valider et compléter SYNLEX. Le rappel et la précision de cette version améliorée de SYNLEX sont ensuite calculés par rapport à un lexique de référence extrait de DICOVALENCE.

Abstract. SYNLEX is a syntactic lexicon extracted semi-automatically from the LADL tables. Like the other syntactic lexicons for French which are both available and usable for NLP (LEFFF, DICOVALENCE), it is incomplete and its recall and precision wrt a gold standard are unknown. We present an approach which goes some way towards adressing these shortcomings. The approach draws on methods used for the automatic acquisition of syntactic lexicons. First, a new syntactic lexicon is acquired from an 82 million words corpus. This lexicon is then used to validate and extend SYNLEX. Finally, the recall and precision of the extended version of SYNLEX is computed based on a gold standard extracted from DICOVALENCE.

Mots-clés : lexique syntaxique, évaluation.

Keywords: syntactic lexicon, evaluation.

1 Introduction

Un lexique syntaxique décrit les propriétés syntaxiques des mots d'une langue. En particulier, un lexique syntaxique associe à chaque foncteur syntaxique un *cadre de sous-catégorisation* spécifiant le nombre et le type (catégorie syntaxique, marqueur introductif, mode, etc.) de ses arguments.

Comme l'ont montré (Carroll & Fang, 2004), un lexique syntaxique exhaustif et détaillé permet d'améliorer les performances des analyseurs syntaxiques. Un tel lexique est également une composante essentielle de tout réalisateur de surface puisqu'il permet de réaliser un contenu

sémantique donné par une phrase bien formée et en particulier, une phrase où chaque foncteur syntaxique a le nombre et le type d'arguments requis par son régime. Plus généralement, un lexique syntaxique est une composante de base pour tout système faisant intervenir soit l'analyse, soit la réalisation.

Pour le francais, il existe à l'heure actuelle trois lexiques syntaxiques disponibles librement et utilisables par des systèmes de traitement automatique des langues : Proton récemment renommé DicoValence, (van den Eynde & Mertens, 2003), Lefff (Clément *et al.*, 2004) et SYNLEX (Gardent *et al.*, 2006). Néanmoins aucun de ces lexiques n'est entièrement satisfaisant pour deux raisons.

Premièrement, aucun de ces lexiques ne couvre l'ensemble des verbes du français. Ainsi pour 8 790 verbes identifiés pour le français dans Morphalou (Romary *et al.*, 2004), DicoValence inclut 3 700 verbes, Lefff 6 798 et SYNLEX 5244.

Deuxièmement, la qualité de leur contenu et plus précisément, leur rappel et leur précision restent inconnus : Pour l'ensemble des entrées contenues dans chacun de ses dictionnaires, on ne connait ni quelle proportion des entrées correctes est présente (rappel) ni quelle est la proportion d'entrées incorrectes (précision).

Dans cet article, nous considérons SYNLEX et présentons une approche qui vise à pallier ces lacunes. L'approche s'appuie sur les méthodes mises au point en acquisition automatique de lexique. Un lexique syntaxique (CORLEX) distinct de SYNLEX est acquis à partir d'un corpus de 82 millions de mots. Ce lexique est ensuite utilisé pour valider et compléter SYNLEX. Le rappel et la précision de SYNLEX, de la version améliorée de SYNLEX et de CORLEX sont ensuite calculés par rapport à un lexique de référence extrait de DICOVALENCE.

L'article est structuré comme suit. La section 2 décrit le processus de création de SYNLEX et présente son format et son contenu. La section 3 présente les travaux visant à valider et à étendre SYNLEX puis commente les résultats obtenus. La section 4 conclut en indiquant les directions de recherche futures.

2 Synlex

Synlex est un lexique créé à partir des tables du LADL (Gross, 1975; Guillet & Leclère, 1992; Boons *et al.*, 1976). Le processus de création a été décrit dans (Gardent *et al.*, 2005b; Gardent *et al.*, 2006; Gardent *et al.*, 2005a) et peut être résumé comme suit :

1. une représentation du contenu des colonnes des tables et de leurs interdépendance est créée manuellement sous la forme d'un graphe *et/ou* dont les noeuds contiennent à la fois des conditions et des pointeurs vers le contenu des colonnes
2. ce graphe *et/ou* est ensuite utilisé en conjonction avec les tables pour produire de facon automatique un lexique syntaxique représentant leur contenu
3. ce lexique est ensuite simplifié pour ne contenir que le type d'information habituellement présente dans un lexique syntaxique (i.e., nombre et types de syntagmes sous-catégorisés par les verbes)

Le format des entrées de SYNLEX est spécifié dans la figure 1 et peut être décrit comme suit. Une entrée se compose d'un verbe, d'une liste d'arguments syntaxiques ayant un rôle séman-

tique, d'une liste optionnelle d'*associés* c-à-d, d'arguments régis par le verbe mais ne remplissant pas de rôle sémantique (e.g., l'explétive *il* dans *il pleut*) et d'une liste de *macros* donnant des informations supplémentaires sur les propriétés syntaxiques du verbes (e.g., contrôle, passivisation). Les *associés* et les *macros* sont des listes finies d'atomes. Un argument en revanche est défini par un triplet de la forme *F :M-C* où F est une fonction grammaticale, M un marqueur optionnel (une préposition ou un clitique indiquant la cliticisation d'un argument en cas d'ambiguité comme par exemple les arguments en *à* qui peuvent se cliticiser soit en y, soit en *lui*) et C est une catégorie syntaxique.

$$
\begin{array}{rcll}
\textit{Entree} & ::= & \textit{Verb} : \langle Arg^{+} \rangle, \textit{Associe}^{*}, \textit{Macro}^{*} & (1) \\
\textit{Arg} & ::= & \textit{Fonction} : \textit{Marqueur} - \textit{Categorie} & (2) \\
\textit{Fonction} & ::= & \textit{suj} \mid \textit{obj} \mid \textit{obja} \mid \textit{objde} \mid \textit{obl} \mid \textit{attr} & (3) \\
\textit{Marqueur} & ::= & \textit{Prep} \mid \textit{Clitic} \mid \textit{Compl} & (4) \\
\textit{Categorie} & ::= & \textit{sn} \mid \textit{pinf} \mid \textit{pcompl} \mid \textit{qcompl} & (5) \\
\textit{Associe} & ::= & \textit{ilimp} \mid \textit{cln} \mid \textit{cla} \mid \textit{cld} \mid \textit{clg} \mid \textit{pron} & (6) \\
\textit{Macro} & ::= & \textit{CtrlArgXArgY} \mid \textit{passivable} \mid \textit{nonPassivable} & (7)
\end{array}
$$

FIG. 1 – SynLex Format

Seules 60% des tables du LADL étant disponibles, nous avons complété manuellement le lexique extrait des tables disponibles avec environ 2 000 verbes et leurs cadres de base. Le lexique SYNLEX résultant contient 5244 verbes et 19127 entrées (paires verbe - cadre) faisant intervenir 726 cadres de sous-catégorisation en considérant les associés et 538 cadres de sous-catégorisation sans associés.

3 Evaluation

Comme nous l'avons mentionné, SYNLEX est produit à partir des tables du LADL par un processus de conversion faisant intervenir une représentation intermédiaire. Or l'information contenue dans les tables peut être inexacte et la conversion dans le format SYNLEX peut introduire des erreurs. Enfin, le lexique produit ne couvre ni l'ensemble des verbes du français, ni nécessairement, l'ensemble des entrées d'un verbe. Il est donc nécessaire à la fois de valider et de compléter le lexique obtenu.

Au cours des 15 dernières années, des travaux (Brent, 1991; Briscoe & Carroll, 1997; Manning, 1993) ont montré qu'il est possible d'extraire un lexique syntaxique d'un corpus en utilisant d'abord un analyseur puis un filtre statistique. L'idée est la suivante. Dans un premier temps, un analyseur déterministe est utilisé pour produire à partir d'un corpus des hypothèses sur les cadres de sous-catégorisation des verbes présents dans ce corpus. Plus précisément, l'analyse produite pour chaque proposition par l'analyseur est utilisée pour associer au verbe de la proposition une description des syntagmes maximaux (groupe nominal, groupe prépositionnel, proposition infinitive, etc.) apparaissant avec ce verbe. Dans un deuxième temps, les hypothèses sont soumises à un calcul statistique et seules sont conservées les hypothèses pour lesquelles la probabilité d'erreur est suffisament basse. Le lexique ainsi obtenu est ensuite évalué (rappel et précision) par rapport à un lexique de référence validé manuellement.

Nous utilisons ici les idées issues de ces travaux pour évaluer la qualité de SYNLEX. D'une part, nous montrons comment un lexique extrait d'un corpus (CORLEX) peut être utilisé pour valider SYNLEX et l'enrichir. Le lexique résultant est appelé XSYNLEX. D'autre part, nous comparons les trois lexiques ainsi créés (SYNLEX, XSYNLEX et CORLEX) avec un corpus de référence (REFLEX) extrait de DICOVALENCE.

3.1 Comparaison et fusion avec un lexique acquis à partir de corpus

Afin d'évaluer la précision et la couverture de SYNLEX, nous commençons par le comparer avec un lexique acquis automatiquement à partir d'un corpus. Ce lexique (CORLEX) est acquis selon la méthodologie décrite ci-dessus : un corpus et un analyseur sont d'abord utilisés pour émettre des hypothèses sur les entrées lexicales (association verbe - cadre) possibles. Ensuite, ces hypothèses sont soumises à un calcul statistique permettant de classifier les hypothèses en hypothèses plausibles et hypothèses non plausibles. Dans ce qui suit, nous détaillons chacun de ces procédés.

Création des hypothèses. Le corpus exploité est un corpus de 82 millions de mots avec 65% d'articles de presse, 30 % de compte rendus de débats parlementaires et 5% de textes littéraires.

L'analyseur (TAGPARSER) est un analyseur robuste ascendant qui exploite des connaissances très fines sur la combinaison des mots grammaticaux classifiés en 300 classes de mots simples ou composés (Francopoulo, 2005). Dans la version actuelle (version 1), mise à part, une catégorisation binaire des adjectifs et l'indication comme quoi le verbe accepte ou non, une complétive, l'analyseur n'utilise pas d'information portant sur la sous-catégorisation des verbes et des noms prédicatifs. La technologie mise en oeuvre combine un automate et une matrice statistique induite à partir d'un corpus de 77 000 mots annotés en syntaxe de surface.

Enfin notons que pour cette première expérience, nous nous sommes limités aux cadres qui sont relativement faciles à détecter pendant l'analyse syntaxique i.e., les cadres ne faisant intervenir ni la fonction oblique, ni la fonction attribut. En outre, les associés (e.g., reflexif intrinsèque, clitique figé) et les macros qui concernent des propriétés syntaxiques non détectables par un analyseur (e.g.,phénomènes de contrôle, acceptation ou non pour les verbes transitifs de la forme passive, etc.) ne sont pas pris en compte.

L'analyse du corpus par TAGPARSER permet d'extraire 38 550 hypothèses où chaque hypothèse est l'association d'un verbe, d'un cadre et d'une fréquence d'apparition de cette association dans le corpus.

Filtrage des hypothèses. Afin d'évaluer la plausibilité des hypothèses émises, nous utilisons un test souvent mis en oeuvre (Brent, 1991; Briscoe & Carroll, 1997; Manning, 1993) par les approches portant sur l'acquisition automatique de lexiques à savoir le test binomial sur les hypothèses (BHT). Ce test calcule la probabilité que m occurrences du cadre c apparaissent avec un verbe v n'acceptant pas ce cadre, étant donné n occurrence de ce verbe. Plus la probabilité est basse, plus l'hypothèse est douteuse et par conséquent, plus il est probable que c est un cadre valide de v.

En pratique, nous fixons à 0.05% le seuil utilisé pour déterminer si ou non une association verbe-cadre apparait suffisament peu fréquemment pour être une erreur. En d'autres termes, toutes les

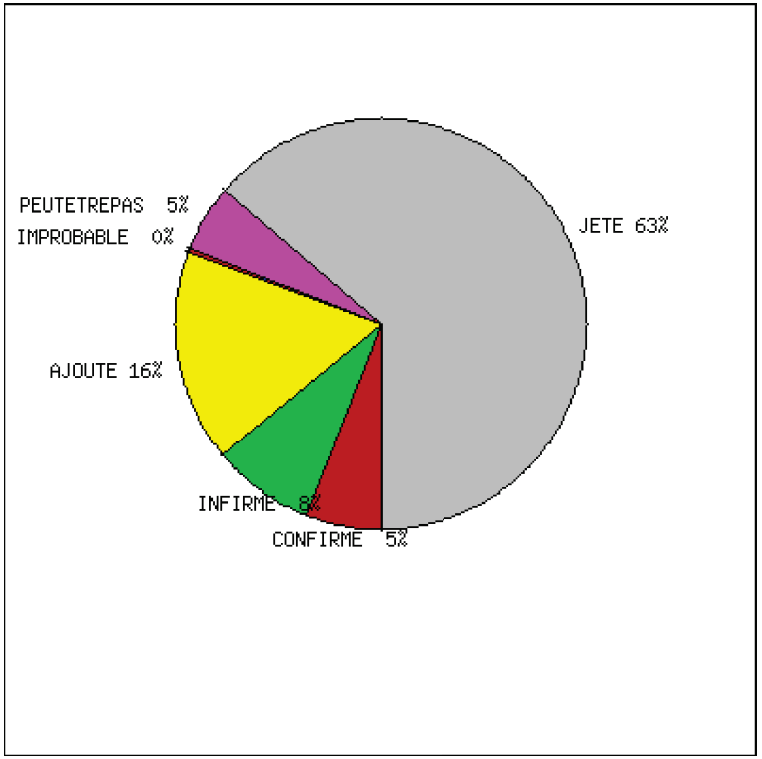


FIG. 2 – Résultats

hypothèses pour lesquelles la probabilité d'erreur donnée par le test BHT est en dessous de 0.05% sont acceptées comme valides – les autres sont rejetées. Pour calculer la probabilité d'erreur des hypothèses émises, nous utilisons le UCS toolkit (`http://www.collocations.de/`). Après filtrage, le lexique syntaxique obtenu (CORLEX) comporte 8 742 entrées.

Comparaison et fusion des deux lexiques (SYNLEX et CORLEX). La figure 2 donne une analyse détaillée des résultats obtenus à partir de l'analyse de corpus. Plus généralement, on peut diviser et classifier les données suivant les critères suivants[1] :

CONFIRMÉ : les entrées présentes dans SYNLEX et dans CORLEX et pour lesquelles la probabilité d'erreur est inférieure à 0.05% .

INFIRMÉ : les entrées présentes dans SYNLEX et dans CORLEX et pour lesquelles la probabilité d'erreur est supérieure à 0.05% .

AJOUTÉ : les entrées absentes dans SYNLEX qui sont présentes dans CORLEX et pour lesquelles la probabilité d'erreur est inférieure à 0.05% .

[1] Les pourcentages sont donnés par rapport à l'union de CORLEX et SYNLEX.

JETÉ : les entrées absentes dans SYNLEX qui sont présentes dans CORLEX et pour lesquelles la probabilité d'erreur est supérieure à 0.05% .

IMPROBABLE : les entrées présentes dans SYNLEX absentes dans CORLEX et pour lesquels le verbe impliqué apparait plus de 5 000 fois dans le corpus.

PEUTETREPAS : les entrées présentes dans SYNLEX absentes dans CORLEX et pour lesquels le verbe impliqué apparait moins de 5 000 fois dans le corpus.

La classe CONFIRMÉ permet de valider la partie de SYNLEX trouvée en corpus et validée par les statistiques. Inversement, la classe INFIRMÉ permet de détecter les entrées de SYNLEX qui sont sans doute incorrectes. Les données montrent en particulier, que sur la base de cette analyse, plus de la moitié des entrées de SYNLEX sont jugées incorrectes.

Par ailleurs, la classe AJOUTÉ permet d'étendre SYNLEX avec les entrées jugées fiables par l'analyse de corpus mais non contenues par SYNLEX. Ceci permet d'augmenter le nombre d'entrées de SYNLEX de 34.56%.

Enfin, les classes IMPROBABLE et PEUTETREPAS regroupent les entrées de SYNLEX qui n'apparaissent pas dans les données extraites du corpus. Les IMPROBABLE sont des cas où le verbe considéré apparaît plus de 5 000 fois dans le corpus mais jamais avec le cadre prescrit par SYNLEX. Ils sont éliminés de SYNLEX. Si le verbe apparait moins de 5 000 fois dans le corpus, l'entrée est conservée mais étiquettée comme peu fiable (PEUTETREPAS).

En résumé, la fusion XSYNLEX de SYNLEX avec CORLEX peut être définie par l'union de CONFIRMÉ avec AJOUTÉ :

$$\text{XSYNLEX} = \text{CONFIRMÉ} \cup \text{AJOUTÉ} \cup \text{PEUTETREPAS}$$

Cependant, cette fusion ne garantit pas un lexique parfait. En effet, la validation statistique reste imparfaite. Par exemple, les meilleurs lexiques extraits pour l'anglais avec des méthodes similaires à celle utilisée ici ont une F-mesure maximum tournant autour de 80 % . La deuxième étape a donc consisté à évaluer les différents lexiques (SYNLEX, CORLEX et XSYNLEX) en mesurant leur rappel et précision par rapport à un lexique de référence REFLEX. L'objectif est de déterminer si l'extension de SYNLEX par les données issues de CORLEX accroit non seulement le nombre d'entrées mais également la qualité du lexique résultant.

3.2 Évaluation de SYNLEX sur un lexique de référence

Une façon de déterminer la qualité d'un lexique consiste à calculer son rappel et sa précision par rapport à un lexique de référence. Soit $Acquis$ le contenu du lexique à évaluer et Ref celui du lexique de référence, précision et rappel sont définis de la façon suivante :

Précision

$$P = \frac{Acquis \cap Ref}{Acquis}$$

La précision indique la proportion d'entrées correctes dans le lexique acquis (combien d'entrées sont correctes ?)

Rappel

$$R = \frac{Acquis \cap Ref}{Ref}$$

Le rappel indique la proportion entre entrées correctes présentes dans le lexique acquis et entrées présentes dans le lexique de référence (combien d'entrées correctes ont été trouvées ?).

Calcul du rappel et de la précision. Pour l'évaluation, nous avons sélectionné 100 verbes présents dans tous les lexiques (i.e., SYNLEX, XSYNLEX, DICOVALENCE et CORLEX) et distribués de façon régulière sur l'échelle du nombre d'apparition dans le corpus.

Pour chacun de ces 100 verbes, nous avons créé un lexique de référence REFLEX à partir de DICOVALENCE. Les entrées de ces verbes ont été épurées des entrées non prises en compte dans CORLEX (c-à-d, les entrées faisant intervenir des arguments obliques ou attributifs) puis traduites dans le format SYNLEX (cf. Figure 1) afin de permettre une comparaison automatique avec SYNLEX, XSYNLEX et CORLEX.

Les performances des statistiques ont été évaluées sur ces 100 verbes à travers quatre expériences visant à mesurer l'impact de la fréquence d'un cadre sur ces performances.

Etant donné C le nombre total d'entrées présentes dans CORLEX, la fréquence f_c d'un cadre c est dite HAUTE si c apparait dans plus de 1% des entrées de CORLEX ($f_c \geq 0.01 \times C$) ; MOYENNE si $0.001 \times C \leq f_c \leq 0.01 \times C$; et BASSE si $f_c \leq 0.0001 \times C$.

Pour chaque lexique (SYNLEX, XSYNLEX et REFLEX), quatre (sous-)lexiques sont créés : un premier contenant toutes les entrées du lexique (TOUT) et trois autres contenant uniquement les entrées faisant intervenir des cadres de haute (HF), moyenne (MF) et basse (BF) fréquence. La référence minimum (baseline) est fixée comme étant le lexique acquis à partir du corpus sans filtrage statistique (toutes les entrées trouvées par TAGPARSER sont prises en compte).

Le rappel et la précision pour chacun des 5 cas considérés sont donnés dans la Figure 3.

Discussion. Ces premiers résultats montrent que pour l'échantillon de cadres considérés (les cadres ne faisant pas intervenir d'obliques ou d'attributs), la couverture et la précision de SYNLEX sont relativement bas. La couverture faible n'est pas surprenante et s'explique du fait de l'incomplétude inhérente aux tables du LADL puisque seules 60% des tables sont disponibles.

La mauvaise précision est en revanche plus surprenante mais peut, peut être, être expliquée par la relative permissivité des tables du LADL : si une construction est possible pour un verbe donné, elle sera marquée comme telle même si elle est très rare.

Un autre facteur contribuant à diminuer la précision concerne la décision de ne pas prendre en compte les associés c-à-d, les arguments régis par le verbe mais ne remplissant pas de rôle sémantique. Or parmi ces associés, on trouve le clitique réfléchi intrinsèque (e.g., *se* dans *s'évanouir*). En conséquence, toutes les entrées faisant intervenir un clitique intrinsèque (l'associé CLR) sont traitées de façon incorrecte comme des entrées sans ce clitique.

Malgré tout, un examen plus approfondi des cas fautifs reste à faire pour déterminer les causes précises de ce manque de précision et éventuellement, y remédier.

Le rappel et la précision de XSYNLEX , le lexique enrichi à partir du corpus, sont relativement

		TOUT	HF	MF	LF
SYNLEX	P	0.30	0.63	0.16	0.02
	R	0.44	0.45	0.47	0.3
	F	0.37	0.54	0.31	0.16
XSYNLEX	P	0.58	0.69	0.23	0.29
	R	0.63	0.66	0.56	0.5
	F	0.59	0.67	0.4	0.4
XSYNLEX + INFIRMÉ	P	0.49	0.61	0.21	0.27
	R	0.76	0.78	0.67	0.5
	F	0.62	0.70	0.44	0.38
BASELINE	P	0.22	0.29	0.07	0.15
	R	0.89	0.95	0.70	0.5
	F	0.56	0.62	0.39	0.32

FIG. 3 – Précision et rappel

bas mais proches de certains résultats obtenus dans la litérature pour des langues autres que l'anglais. (Fast & Przepiórkowski, 2005) par exemple, cite un rappel de 47% et une précision de 49% pour une expérience similaire sur le polonais. Pour ce lexique, le rappel et la précision sont meilleurs que pour SYNLEX. En d'autres termes, le lexique extrait du corpus permet de valider et d'étendre la partie de SYNLEX faisant intervenir les cadres considérés pour l'acquisition automatique.

Enfin, les données concernant XSYNLEX+ INFIRMÉ montrent qu'ignorer la plausibilité statistique des hypothèses (i.e., conserver les entrées de SYNLEX qui sont infirmées par les statistiques) permet d'améliorer le rappel (0.76 contre 0.63 dans XSYNLEX) au détriment bien sûr de la précision (0.49 contre 0.58 dans XSYNLEX).

4 Conclusion et perspectives

Comme nous l'avons mentionné dans l'introduction, trois lexiques syntaxiques sont actuellement disponibles et utilisables dans le domaine du traitement automatique des langues. Cependant, ils sont tous incomplets et leur contenu n'a pas fait l'objet d'une évaluation permettant de déterminer rappel et précision.

Le travail présenté dans cet article est un premier pas vers la définition d'une procédure d'évaluation et de fusion de ces lexiques.

Il montre en particulier que DICOVALENCE peut servir de base à la création d'un lexique de référence permettant ainsi de calculer le rappel et la précision de lexiques créés de façon automatique ou semi-automatique.

Il montre également, qu'un lexique acquis à partir d'un corpus peut permettre d'améliorer la

couverture et la précision d'un lexique existant ; et plus généralement, que la comparaison et la fusion de plusieurs lexiques pourrait permettre à relativement court terme de produire un lexique syntaxique du français complet et de bonne qualité.

Néanmoins, plusieurs aspects méritent d'être approfondis.

Tout d'abord, notons que l'évaluation de SYNLEX présentée ici est très partielle puisqu'elle ne porte que sur 33 des 726 cadres présents dans SYNLEX. Une évaluation plus extensive prenant en compte les obliques et les attributs est donc nécessaire.

Un second point concerne la procédure d'acquisition automatique. En effet, l'approche présentée ici est une approche préliminaire qui peut être améliorée sur au moins deux points à savoir, la qualité des hypothèses émises d'une part et la qualité du filtre statistique d'autre part.

Les hypothèses émises peuvent être affinées par l'emploi d'un analyseur plus performant – par exemple, en utilisant une information de sous-catégorisation pour informer l'analyseur ou encore en utilisant un analyseur profond plutôt que local. Une autre possibilité que nous entendons explorer prochainement, est d'utiliser plusieurs analyseurs en parallèle et de comparer/fusionner leurs résultats par un système de vote.

Les travaux fait sur l'anglais suggèrent en outre que le filtre statistique peut être amélioré de deux façons. Ainsi (Briscoe & Carroll, 1997) montre que le seuil permettant de déterminer l'acceptabilité d'une hypothèse doit être fixé différemment suivant le type de cadre considéré plutôt que de façon uniforme pour l'ensemble des hypothèses comme nous l'avons fait ici. Et (Korhonen, 2002) montre que l'utilisation de techniques de lissages informées par les classes sémantiques de verbes permet d'améliorer les résultats. L'exploitation de ces résultats devrait permettre d'améliorer la qualité du lexique extrait.

Une troisième point, plus ouvert celui-là, concerne l'élargissement des méthodes explorées à l'ensemble du lexique et en particulier au traitement des macros. Comme nous l'avons vu, SYNLEX, LEFFF et DICOVALENCE contiennent outre des informations portant sur la valence (arguments régis par le verbe remplissant ou non un rôle sémantique), des informations portant sur les phénomènes de contrôle, la passivation, la possibilité pour un verbe d'être utilisé dans une tournure impersonnelle, etc. Si elles sont utiles pour le traitement automatique des langues et en particulier, pour l'analyse et la réalisation de surface, ces informations ne peuvent pas être extraites à partir des corpus par les techniques utilisées en acquisition automatique de lexique. Elles sont en revanche partiellement présentes dans les lexiques existants (LEFFF, DICOVALENCE et SYNLEX). Une question intéressante est donc de savoir comment cette information peut être utilisée pour informer la complétion d'un lexique partiellement sous-spécifié dans cette dimension. Ou en d'autres termes, comment un lexique acquis à partir de corpus peut être fusionné avec un ou des lexiques acquis par des méthodes «symboliques» (LEFFF, SYNLEX) de façon à enrichir la partie acquise statistiquement avec l'information additionnelle contenue dans les lexiques symboliques.

Dans tous les cas, la précision relativement basse des lexiques produits suggère qu'une phase de validation manuelle est nécessaire. Dans cette optique, une approche qui consiste à privilégier (dans une juste mesure) le rappel plutôt que la précision est sans doute préférable (il est plus facile d'éliminer que d'ajouter). Ce qui suggère en particulier, que XSYNLEX+ INFIRMÉ est préférable à XSYNLEX et plus spécifiquement, que l'extraction de SYNLEX à partir des tables est utile.

Références

BOONS J.-P., GUILLET A. & LECLÈRE C. (1976). *La structure des phrases simples en français. I : Constructions intransitives*. Droz, Genève.

BRENT M. (1991). Automatic acquisition of subcategorisation frames from untagged text. In *Proceedings of the 29th Meeting of the ACL*, p. 209–214, Berkeley.

BRISCOE T. & CARROLL J. (1997). Automatic extraction of subcategorisation from corpora. In *Proceedings of the 5th ANLP conference*, p. 356–363.

CARROLL J. & FANG A. (2004). The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, p. 107–114, Sanya City, China.

CLÉMENT L., SAGOT B. & LANG B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of LREC'04*, Lisbonne.

FAST J. & PRZEPIÓRKOWSKI A. (2005). Automatic extraction of polish verb subcategorisation. an evaluation of common statistics. In *Proceedings of the 2nd Language and Technology conference*, p. 191–195.

FRANCOPOULO G. (2005). Tagparser et technolangue-easy. In *Actes de l'atelier Easy, TALN*.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2005a). Extracting subcategorisation information from Maurice Gross' Grammar Lexicon. *Archives of Control Sciences*, **15**(LI), 253–264.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2005b). Maurice gross' grammar lexicon and natural language processing. In *Proceedings of the 2nd Language and Technology Conference*, Poznan, Poland.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2006). Extraction d'information de sous-catégorisation à partir des tables du ladl. In *Actes de La 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*.

GROSS M. (1975). *Méthodes en syntaxe*. Hermann.

GUILLET A. & LECLÈRE C. (1992). *La structure des phrases simples en français. Constructions transitives locatives*. Droz, Genève.

KORHONEN A. (2002). *Subcategorization Acquisition*. PhD thesis, University of Cambridge.

MANNING C. (1993). Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st Meeting of the ACL*.

ROMARY L., SALMON-ALT S. & FRANCOPOULO G. (2004). Standards going concrete : from LMF to morphalou. In *Workshop on Electronic Dictionaries*, Geneva, Switzerland.

VAN DEN EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies 13, 63-104*.

Session
Morphologie

Analyse automatique *vs* analyse interactive : un cercle vertueux pour la voyellation, l'étiquetage et la lemmatisation de l'arabe

Fathi DEBILI[1], Zied BEN TAHAR[1], Emna SOUISSI[2]
[1] LLACAN, INALCO, CNRS
7, rue Guy Môquet, 94801 Villejuif cedex, France
[2] ESSTT, 5, Avenue Taha Hussein – 1008 Tunis
fathi.debili@wanadoo.fr, bentaharzied@gmail.com,
emna.souissi@planet.tn

Résumé. Comment produire de façon massive des textes annotés dans des conditions d'efficacité, de reproductibilité et de coût optimales ? Plutôt que de corriger les sorties d'analyse automatique moyennant des outils d'éditions éventuellement dédiés, ainsi qu'il est communément préconisé, nous proposons de recourir à des outils d'analyse interactive où la correction manuelle est au fur et à mesure prise en compte par l'analyse automatique. Posant le problème de l'évaluation de ces outils interactifs et du rendement de leur ergonomie linguistique, et proposant pour cela une métrique fondée sur le calcul du coût qu'exigent ces corrections exprimé en nombre de manipulations (frappe au clavier, clic de souris, etc.), nous montrons, au travers d'un protocole expérimental simple orienté vers la voyellation, l'étiquetage et la lemmatisation de l'arabe, que paradoxalement, les meilleures performances interactives d'un système ne sont pas toujours corrélées à ses meilleures performances automatiques. Autrement dit, que le comportement linguistique automatique le plus performant n'est pas toujours celui qui assure, dès lors qu'il y a contributions manuelles, le meilleur rendement interactif.

Abstract. How can we massively produce annotated texts, with optimal efficiency, reproducibility and cost? Rather than correcting the output of automatic analysis by means of possibly dedicated tools, as is currently suggested, we find it more advisable to use interactive tools for analysis, where manual editing is fed in real time into automatic analysis. We address the issue of evaluating these tools, along with their performance in terms of linguistic ergonomy, and propose a metric for calculating the cost of editing as a number of keystrokes and mouse clicks. We show, by way of a simple protocol addressing Arabic vowellation, tagging and lemmatization, that, surprisingly, the best interactive performance of a system is not always correlated to its best automatic performance. In other words, the most performing automatic linguistic behavior of a system is not always yielding the best interactive behavior, when manual editing is involved.

Mots-clés : analyse automatique *vs* interactive ; annotation séquentielle, parallèle ; voyellation, lemmatisation, étiquetage de l'arabe ; métrique pour l'évaluation de l'analyse interactive.

Keywords: automatic versus interactive analysis of Arabic, proposal of metrics for evaluating the interactive analysis, design and implementation of software for interactive vowellation, lemmatisation and POS-tagging of Arabic, evaluation.

1 Introduction

L'analyse automatique semble avoir précédé l'analyse interactive, laquelle signifie intervention manuelle. Elle a été la préoccupation première des chercheurs, pour la plupart d'entre eux et dès le départ[1], et sans doute restera-t-elle longtemps encore le but à atteindre, la dimension à parfaire. L'analyse manuelle que nous dirons « artisanale » a, elle aussi, été pratiquée d'emblée, avec des objectifs divers, en particulier celui de la confection de corpus annotés orientés vers l'apprentissage ou l'évaluation. Même si l'on s'est très vite rendu compte de la difficulté matérielle qu'il y avait à produire de l'analyse manuelle, ce n'est que tardivement, sous la pression d'une double exigence, de performances et de plus large couverture, que l'on y a consacré des efforts soutenus. Avec la rédaction de guides d'annotation pour rendre l'opération autant que faire se peut reproductible (cf. l'action GRACE par exemple, Adda et al. 1999, Véronis, 1999, Abeillé et Clément, 2003). Puis avec la confection d'outils informatiques dédiés où la part de l'automatique au service du manuel a été peu à peu introduite et amplifiée (Habert, 2005). Le présent travail s'inscrit dans cette dynamique qu'il prolonge. Nous abordons les problèmes que pose l'annotation massive, manuellement vérifiée et corrigée, de corpus arabes. Autrement dit, de l'analyse morpho-grammaticale interactive de l'arabe.

Au travers des difficultés que présentent la voyellation, l'étiquetage et la lemmatisation de l'arabe, cf. partie 2, des coûts prohibitifs qu'elles engendrent sous l'angle de la vérification et saisie manuelle, cf. partie 3, nous décrivons, partie 4, les spécifications qui nous ont amenés à développer une analyse interactive vue non pas comme indépendante de l'analyse automatique, même si elle en utilise les résultats qu'elle est sensée lui renvoyer éventuellement corrigés, mais bien comme une extension rétroactive de celle-ci.

Soulevant le problème que pose l'évaluation des performances de l'annotation interactive, nous montrons qu'il y a intrication entre les deux processus, automatique et interactif, où le service rendu mutuel va au-delà du simple échange de données annotées. L'on constate en effet que l'exigence de meilleures performances pour les procédures d'analyse interactive, qui passent par la définition de diverses ergonomies linguistiques intuitives et efficaces, amène à reconsidérer la conception même des algorithmes de la dimension automatique.

Deux ergonomies linguistiques se dégagent. La première, séquentielle, est celle, classique, qui vient naturellement à l'esprit. Elle est liée au fait que les vérifications annotations manuelles nécessitent en général que soit consulté le contexte du mot en cours de vérification. Cette ergonomie s'avère très lente, et donc peu productive, cf. partie 5. La seconde, parallèle, essaie de parer à cette lenteur en mettant à profit le fait que bon nombre de mots apparaissent souvent, le gain projeté étant alors que l'on puisse tous les vérifier et annoter en même temps. Cette ergonomie s'avère plus productive mais est plus difficile à mettre en œuvre, cf. partie 6.

Faisant ainsi converger nos préoccupations vers la réalisation d'un système intégré, comment évaluer les performances des ergonomies linguistiques et interactives qui en constituent l'interface, performances qui relèvent *a priori* du qualitatif, et qui en même temps restent dépendantes des performances des traitements d'analyse automatique qui, eux, constituent le cœur du système ? Une métrique et un protocole expérimental sont proposés pour la mesure

[1] Les diverses applications *assistées par ordinateur* (*x*. A. O.) ne visaient pas la confection massive de données dictionnairiques ou textuelles annotées.

des performances de l'analyse interactive, lesquelles ne se calculent pas de la même façon que celles de l'analyse automatique.

Résultats d'expérimentations et commentaires sont livrés parties 5 et suivantes.

2 Des niveaux d'ambiguïté élevés

Le mot arabe, tel qu'on le rencontre dans les textes, c'est-à-dire sous sa forme fléchie, simple ou agglutinée (*proclitique+forme simple+enclitique*, que nous conviendrons d'appeler hyper-forme), présente des niveaux d'ambiguïté segmentale, vocalique, casuelle, lemmatique, et grammaticale relativement élevés. Le tableau 1 donne à titre indicatif les valeurs moyennes mesurées en définition (comptages effectués sur des données dictionnairiques : un dictionnaire de 66 millions d'entrées non voyellées obtenues par synthèse lexico-syntagmatique, un autre de 157 mille entrées issues d'un corpus de 2 millions d'occurrences), et en usage (comptages effectués sur des données textuelles : ici, sur les 2 millions d'occurrences du corpus précité).

Ambiguïté	Segmentale	Vocalique et Casuelle	Lemmatique	Grammaticale
Dictionnaire 66.10^6	1,08	2,17	1,68	2,99
Sous lexique 157 031	1,26	6,40	2,65	9,16
En usage	1,32	7,84	3,66	10,76

Tableau 1 : Niveaux d'ambiguïté de l'hyper-forme arabe

Ces valeurs placent l'arabe à des niveaux d'ambiguïté sensiblement plus élevés que ceux du français. Elles se rapportent en effet, ainsi que nous venons de le dire, non pas aux formes simples de l'arabe, dont les niveaux d'ambiguïté sont plus élevés encore (Debili et al. 2002), mais aux formes simples et agglutinées. Une autre mesure, plus globale, a pu être effectuée. Elle se rapporte au niveau d'ambiguïté composée, c'est-à-dire toutes ambiguïtés segmentales, vocaliques, casuelles, lemmatiques, et grammaticales confondues. La synthèse lexico-syntagmatique donne en effet pour 500 mille formes fléchies simples non voyellées arabes, 305 millions de formes simples et agglutinées, voyellées, lemmatisées et étiquetées, différentes, correspondant à 66 millions de formes simples et agglutinées non voyellées. Le rapport de 305 sur 66 conduit à une ambiguïté moyenne d'environ 4,6 acceptions morpho-grammaticales différentes par entrée. Ce chiffre est de 14,7 si les comptages sont effectués sur le sous lexique de 157 mille entrées. En usage, comptages effectués sur le texte de 2 millions d'occurrences, cette moyenne passe à 16,7 acceptions morpho-grammaticales différentes par occurrence. L'on peut remarquer, incidemment, que la répétition textuelle semble ainsi puiser davantage dans l'ambigu que dans le non ambigu.

Ces niveaux d'ambiguïtés sont relativement importants. Nous ne disposons pas de chiffres équivalents pour le français ou l'anglais. Il nous faudrait pour cela considérer les ambiguïtés liées non pas seulement aux formes fléchies, mais aussi aux syntagmes constitués de ces formes simples et des mots vides (articles, prépositions, pronoms, etc.) qui peuvent leurs être adjoints, afin d'établir le parallèle avec l'arabe où ces mots s'attachent sous forme de proclitiques et d'enclitiques. Dans la terminologie de Lucien Tesnière, considérer les *mots constitutifs* accompagnés de leurs *mots subsidiaires* ou *satellites* (Tesnière, 1969, p. 57, §18).

Dans une perspective d'annotation manuelle, au-delà des difficultés à caractère linguistique (définition des étiquettes, critères de choix, etc.) dont nous admettrons qu'elles puissent être comparables d'une langue à une autre, ces niveaux d'ambiguïté indiquent que l'opération d'annotation sera sans doute comparativement plus coûteuse au plan matériel qu'elle ne peut l'être pour le français par exemple, l'étendue des choix étant plus large. Avec la saisie des voyelles, la situation va être plus critique encore.

3 Des coûts d'annotation et de saisie élevés

En effet, en arabe, la plupart des lettres (87% en définition, 77% en usage) demandent pour être voyellées d'être accompagnées d'un signe diacritique dont la saisie coûte 2 frappes au clavier, à l'image du tréma en français. La saisie des lettres voyellées en arabe est donc particulièrement coûteuse : 3 frappes en l'occurrence, soit autant que pour les lettres avec tréma en français. Le tableau 2 donne le coût moyen du caractère exprimé en nombre de frappes, calculé pour différents corpus : français (673 mille mots), anglais (650 mille mots), arabe voyellé (800 mille mots), et arabe non voyellé (2 millions de mots).

	Coût moyen du caractère	Proportion des signes diacritiques	Proportion dans le coût de la saisie
Anglais	1,00001	0,0005	0,001
Français	1,003	3,51	3,84
Arabe non voyellé	1,037	-	-
Arabe voyellé	1,46	43,7%	59,9%

Tableau 2 : Coût moyen du caractère en nombre de frappes

Ces chiffres signifient que la saisie d'un texte de N caractères (lettres avec ou sans signe diacritique) coûtera approximativement N×1,00001 frappes au clavier si le texte est en anglais, contre N×1,003 si le texte est en français, N×1,037 si le texte est en arabe non voyellé, mais N×1,46 si le texte est en arabe voyellé ! Si l'on ajoute que la voyellation d'un texte préalablement saisi ne coûte pas moins, mais autant que de le ressaisir entièrement voyellé (Debili et Fluhr, 2006), alors l'annotation vocalique de l'arabe, sans autre précaution, s'avère prohibitivement coûteuse.

Ces caractérisations sont bien entendu liées à la technologie, aux claviers respectivement associés à chacune des trois langues. Elles offrent une sorte d'évaluation *a posteriori* des standards et normes en vigueur qu'elles sont susceptibles de conforter ou d'infléchir[2]. Mais

[2] En incitant à les amender pour un meilleur rendement. Car sous cet angle, la technologie ne semble pas conférer les mêmes avantages aux langues qu'elle prend en charge. Sur un autre plan, ces comptages et observations suggèrent que les systèmes d'écriture qui persistent ou qui s'installent dans l'usage sont ceux dont le coût est proche de 1, tel que celui de l'anglais, du français, ou de l'arabe non voyellé. On peut remarquer que l'arabe voyellé qui présente un coût de 1,46 le caractère est très peu pratiqué. Même si les raisons qui sous tendent ce constat sont sans doute de nature bien plus complexe, n'y a-t-il pas là un seuil au-delà duquel un système d'écriture n'est plus pratiqué ?

elles permettent aussi, en appréhendant les difficultés que pose la confection massive de corpus annotés sous un angle matériel, d'introduire, aux côtés des métriques d'évaluation des procédures d'analyse automatique classiques, une métrique pour l'évaluation quantitative des processus d'analyse interactive, fondée sur le calcul des coûts qu'engendrent précisément les nécessaires interventions manuelles.

4 Evaluation de l'annotation interactive

Un système d'annotation automatique est performant à 100% lorsque ses résultats sont jugés totalement conformes à une annotation manuelle. Ce critère ne vaut évidemment pas pour un système d'annotation interactif, puisque par définition la conformité est ici atteinte à la fin du processus. Un système d'annotation interactif est en fait d'autant plus performant que le nombre de manipulations imposées à l'annotateur pour accomplir une tâche donnée est petit. Lorsque les performances de sa composante automatique, ici, d'étiquetage, de lemmatisation, et de voyellation sont totales, cet objectif est évidemment atteint puisque pour chaque occurrence de mot, les trois propositions – de lemmatisation, d'étiquetage, et de vocalisation – classées en tête de leurs listes respectives s'avèreront systématiquement correctes. Dans ces conditions les manipulations de l'annotateur se réduisent à de simples validations qui ne lui coûtent en nombre d'opérations qu'une seule action (frappe au clavier, clic de souris, pointage sur un écran tactile, etc.). C'est une situation idéale, mais que l'on ne parvient pas atteindre pour toutes les occurrences qui constituent un corpus, les performances des programmes d'analyse automatique étant, comme on le sait, en deçà du 100%. Pour ces occurrences, le coût de l'annotation est d'autant plus élevé que les solutions proposées par la composante automatique se trouvent situées loin dans les listes des voyellations, des étiquettes et des lemmes résiduels, c'est-à-dire des solutions potentielles qui n'ont pu être éliminées. L'opération d'annotation interactive la plus coûteuse advient lorsque la résolution est en queue de liste, ou plus grave, lorsqu'elle ne s'y trouve pas du tout.

Les performances de l'analyse interactive dépendent des performances de l'analyse automatique, mais tandis que dans un cas, elles sont évaluées au nombre ou à la proportion des occurrences qui sont correctement annotées ou non, elles sont évaluées dans l'autre cas au nombre ou à la proportion des interventions manuelles effectives nécessaires pour valider le correct, et corriger l'incorrect. En cela, et quoique corrélées aux extrêmes, ces deux performances sont complémentaires et ne renseignent pas de la même façon. L'évaluation sous l'angle interactif jette en fait un autre regard sur les performances de la composante automatique, et peut conduire, ainsi que nous allons le montrer, jusqu'à suggérer d'en modifier la conception ou le comportement interne, aboutissant ainsi à des spécifications d'analyseurs automatiques différents, selon qu'ils sont destinés à un usage interactif, ou à un usage automatique pur, du moins si leurs performances restent en deçà d'un certain seuil.

Il y a cercle vertueux parce que les actions manuelles, dès lors qu'elles sont prises en compte, modifient à leur tour de façon dynamique les performances de l'analyse automatique. En effet, en éliminant les ambiguïtés là où elles résistent, ces actions améliorent les performances locales des règles automatiquement mises en jeu, et donc les performances globales de l'analyse automatique, laquelle, offrant de meilleurs résultats, diminue d'autant la charge manuelle, améliorant ainsi les performances de la partie interactive, et ainsi de suite.

Mais l'enseignement qu'apportent l'évaluation interactive et son impact sur la définition de l'analyse automatique va plus loin encore. L'on s'aperçoit que l'ordre d'application des règles qui conduit aux meilleures performances automatiques n'est pas forcément celui qui conduit

aux meilleures performances interactives, sauf cas extrême d'une annotation automatique totalement réussie où les deux performances se rejoignent alors. Ce point nous paraît important. Nous ne pointons pas le fait que, ayant bénéficié d'une contribution humaine externe, alors l'analyse automatique produit de meilleurs résultats. Cela est entendu. Nous disons que *les meilleures performances interactives d'un système ne sont paradoxalement pas toujours corrélées à ses meilleures performances automatiques*. Autrement dit, que le comportement linguistique automatique le plus performant n'est pas toujours celui qui assure, dès lors qu'il y a interférence manuelle, le meilleur rendement interactif. Nous décrivons dans le paragraphe suivant le protocole expérimental et les résultats qui ont conduit à ce constat contre intuitif.

5 Annotation interactive séquentielle

L'ergonomie interactive qui vient en premier à l'esprit est séquentielle. Elle est liée à la nature des ambiguïtés que nous voulons lever, ici les ambiguïtés que pose la voyellation, la lemmatisation, et l'étiquetage de l'arabe, et au fait que pour lever ces ambiguïtés, le recours au contexte s'impose. De sorte que c'est tout naturellement que l'on s'oriente vers une lecture séquentielle lorsque l'on souhaite établir ou vérifier les annotations d'un texte.

Ayant à accomplir pour chaque occurrence trois choix, – de sa voyellation, de son lemme, de son étiquette, – et dans la mesure où ces choix peuvent interférer, c'est-à-dire influer de façon dynamique sur l'ordre selon lequel sont présentées les solutions des annotations non encore fixées, plusieurs (6 au total) séquences ou protocoles d'intervention peuvent être proposés à l'annotateur, selon que l'on commence par l'un ou l'autre de ces trois choix, et que l'on poursuive ainsi. L'arborescence Figure 1 donne les six cas possibles. A ces six séquences ou protocoles, il convient d'ajouter un septième, celui où les choix resteraient indépendants : pas d'interférence ; on ne retient pas que la résolution de l'une des trois valeurs puisse réduire l'ambiguïté qui porte sur les deux autres, puis, en cascade, que la résolution d'une deuxième puisse réduire l'ambiguïté de la dernière.

Deux protocoles sont *a priori* privilégiés : Etiquetage, puis Lemmatisation, puis Voyellation (séquence ELV, à gauche sur la figure 1), et Voyellation, suivie de Lemmatisation, puis Etiquetage (séquence VLE, à droite). Le premier donne l'ordre selon lequel opèrent les traitements automatiques, la machine donc. Le second donne l'ordre selon lequel opèrent préférentiellement les annotateurs, c'est-à-dire selon lequel les traitements manuels sont effectués. Ces deux protocoles sont privilégiés en vertu de considérations qui sont liées à leurs performances attendues d'une façon générale, et supposées être les meilleures par opposition aux performances des autres protocoles.

Dans le premier cas, les meilleures performances d'analyse automatique attendues semblent pouvoir provenir d'une succession Etiquetage, Lemmatisation, puis Voyellation, à l'image par exemple de ce qui est communément retenu pour le français. En effet, dans une approche modulaire, les règles utiles pour lever ces différents types d'ambiguïtés paraissent pouvoir être plus facilement apprises pour le niveau grammatical, que pour les deux autres niveaux. Ce sont donc en premier les ambiguïtés grammaticales qui sont réduites. Les ambiguïtés lemmatiques et vocaliques, pour lesquelles il semble plus difficile ou plus long de rassembler des règles qui leurs soient propres, peuvent néanmoins bénéficier de ces réductions d'ambiguïtés grammaticales : précisément, en écartant les candidats lemmes et/ou voyellations exclusivement liés aux étiquettes éliminées. Par exemple, l'élimination durant la phase d'étiquetage de l'étiquette *nom* permet de ne plus retenir au compte du mot *élève* que le

lemme *élever*. Le lemme *élève* n'étant que *nom*, il est éliminé en même temps ou suite à l'élimination de l'étiquette *nom*.

Dans le second cas, ce sont les performances globales du processus interactif machine-annotateur que l'on essaie de maximiser. Le facteur humain est ici prépondérant. Quel est le protocole ergonomique qui assure la meilleure efficacité, le meilleur rendement ? Il semble raisonnable de supposer que les annotateurs auront plus de facilités à d'abord Voyeller, Lemmatiser, puis Etiqueter (parcours VLE, à droite), ou à Lemmatiser, puis Voyeller, puis Etiqueter, (parcours LVE, au centre), que de commencer par Etiqueter (parcours de gauche).

Ces parcours interactifs induisent des comportements linguistiques machine différents. Du fait que les règles interagissent entre elles, on ne sait pas *a priori* lequel de ces parcours ou comportements est le plus performant sous l'angle automatique, ni lequel est le plus performant sous l'angle interactif.

Pour mesurer ces performances, nous avons imaginé et mis en œuvre le protocole expérimental simple suivant. Partant d'un corpus préalablement annoté de 145 mille hyper-formes (toutes entièrement voyellées, lemmatisées et étiquetées), nous en avons extrait les fréquences relatives : f(étiquette | Mot Non Voyellé) = Nbre(MNV, étiquette)/Nbre(MNV) ;
f(lemme | mot non voyellé) ; f(voyellation | mot non voyellé) ;
f(lemme | mot non voyellé, étiquette) ; etc., voir légende de la figure 1.

MNV

Traitements automatiques — **Traitements interactifs**

	Etiquette	Lemme	Voyellation
A	76,76%	93,04%	84,18%

	Lemme	Voyellation	Etiquette	Voyellation	Etiquette	Lemme
B	74,06%	75,96%	74,04%	80,99%	75,87%	73,78%
C	96,39%	98,91%	79,68%	87,18%	90,68%	96,28%

	Voyellation	Lemme	Voyellation	Etiquette	Lemme	Etiquette
D	73,88%	73,87%	73,86%	73,77%	73,78%	73,77%
E	99,73%	97,19%	99,73%	91,48%	97,19%	91,48%
F	**0,43**	0,43	0,41	**0,36**	0,37	0,37
G	**0,21**	0,21	0,20	**0,17**	0,18	0,18

Ligne **A** : Performances automatiques, Application des règles f(E|MNV), f(L|MNV), f(V|MNV).
Ligne **B** : Performances automatiques, Application des règles f(L|MNV, E), f(V|MNV, E), f(E|MNV, L), f(V|MNV, L), f(E|MNV, V), f(L|MNV, V).
Ligne **C** : Performances interactives, Application des règles f(L|MNV, E), f(V|MNV, E), f(E|MNV, L), f(V|MNV, L), f(E|MNV, V), f(L|MNV, V). Ici, dans les conditions | MNV, *y*), *y* est correct.
Ligne **D** : Performances automatiques, Application des règles f(V|MNV, E, L), f(L|MNV, E, V), f(V|MNV, L, E), f(E|MNV, L, V), f(L|MNV, V, E), f(E|MNV, V, L).
Ligne **E** : Performances interactives, Application des règles f(V|MNV, E, L), f(L|MNV, E, V), f(V|MNV, L, E), f(E|MNV, L, V), f(L|MNV, V, E), f(E|MNV, V, L). Ici, dans | MNV, *y*, *z*), *y* et *z* sont corrects.
Ligne **F** : Coût ergonomique des interventions manuelles, annotation séquentielle} *exprimé en nombre moyen*
Ligne **G** : Coût ergonomique des interventions manuelles, annotation parallèle } *de frappes ou clics par mot.*
MNV : hyper-forme non voyellée ; E : étiquette grammaticale ; V : voyellation ; L : lemme

Figure 1 : Performances des analyses automatique et interactive liées aux six séquences possibles d'application des règles et d'interventions manuelles

Utilisant ces fréquences comme autant de règles unaires que nous avons réappliquées en cascade le long des six parcours, nous en avons calculé les performances, et de façon rétrospective les coûts qu'auraient engendrés les interventions manuelles nécessaires pour en corriger les écarts.

La figure 1 liste ces résultats. Les lignes A, B, et D donnent les performances de l'étiquetage, lemmatisation et voyellation automatiques mettant en œuvre les séquences de règles f(x|MNV), f(x|MNV, y), et f(x|MNV, y, z), avec, selon les parcours, x, y, z = E, L ou V. Les lignes C et E donnent les performances issues de l'application de ces mêmes règles, mais avec y et z manuellement corrigées. La ligne F donne les coûts moyens rapportés au mot, selon les parcours, des diverses interventions manuelles, interventions qui consistent à simplement valider si les choix machine sont corrects (coût nul), et à désigner au moyen de la souris les valeurs potentielles E, L, et V qui conviennent étant données l'occurrence MNV et son contexte, si celles-ci ne sont pas proposées en première position dans leurs listes respectives. Le coût partiel est nul si la valeur E, L, ou V classée première par le système est correcte. Il est sinon d'autant plus élevé que la résolution est classée loin dans la liste des solutions potentielles non éliminées. La formule retenue pour calculer le coût global d'une opération d'annotation interactive est simple :

$$\text{Coût d'annotation séquentielle} = \Sigma_{i=1 \text{ à nombre de mots du corpus}} \Sigma_{x=\text{E, L, ou V}} (\text{rang de la résolution } x_i - 1)$$

Nous constatons que les coûts d'annotation rapportés au mot sont tous différents. Mais surtout que coûts d'annotation interactive et performances d'analyse automatique ne sont pas corrélés, comme l'on aurait pu s'attendre. Le coût le plus bas (42 652 clics ou déplacements de curseur, soit 0,36 clic ou frappe en moyenne par mot, séquence LVE) ne correspond pas au comportement automatique le plus performant (73,88% des mots tous correctement annotés, séquence ELV) qui, lui, réclame 50590 interventions manuelles pour en corriger les écarts, soit 0,43 clic en moyenne par mot. Les séquences d'annotation qui donnent les coûts les plus bas s'avèrent être les séquences qui consistent à commencer par la vérification de la lemmatisation ou voyellation, puis respectivement la lemmatisation ou voyellation, puis étiquetage, tandis que les séquences qui donnent les meilleures performances automatiques s'avèrent être celles qui commencent par l'étiquetage. Les premières correspondent aux trois parcours dessinés à droite sur la figure 1, les secondes, au trois parcours de gauche. Cette distribution spatiale qui partage la figure 1 en deux parties selon les niveaux de performances et de coûts (voir fig. 2), révèle qu'il ne faut pas privilégier un seul comportement automatique, le plus performant en l'occurrence. D'autres comportements moins performants peuvent se révéler meilleurs dès lors qu'il y a interaction. Elle confirme aussi le bien fondé des approches *a priori* préconisées, selon qu'elles sont orientées vers l'autonomie, ou vers l'interactivité.

6 Annotation interactive parallèle

Si les hapax sont rares (5 à 12% selon les corpus dont nous disposons), et les proportions des mots qui apparaissent deux fois ou plus dans un corpus, importantes, ne pourrait-on factoriser les annotations, c'est-à-dire voyeller, lemmatiser et étiqueter en même temps toutes les occurrences d'un même mot ? Car outre les gains de productivité attendus, ces conditions pourraient aussi assurer une meilleure reproductibilité dans la mesure où, opérant de façon contrastive (toutes les occurrences en contexte d'un même mot sont visibles en même temps), l'annotateur pourrait en effet décider de façon plus homogène. Ces considérations nous ont amené à dessiner les contours d'une ergonomie d'annotation parallèle dont la figure 1, ligne G, donne, pour le même corpus, les performances calculées de façon rétrospective en se fondant sur les mêmes conventions de coût.

La comparaison des lignes F et G révèle un gain de facteur 2 : l'annotation parallèle coûte approximativement deux fois moins cher que l'annotation séquentielle. Mais l'on constate surtout que les observations que nous avons pu faire plus haut restent vraies. Avec une acuité légèrement accrue, nous remarquons en effet que l'annotation parallèle la moins coûteuse n'est pas corrélée au traitement automatique le plus performant. Et que la partition droite gauche observée plus haut, selon les niveaux de performance ou de coût, est confortée.

Dans cette ergonomie, l'annotation interactive n'est plus appliquée à toutes les occurrences du corpus prises une à une, mais aux seules entrées du lexique qui leur correspond. Dans le cas présent, aux seules 24291 différentes hyper-formes non voyellées qui constituent le lexique du corpus, et non aux 117900 occurrences reconnues de ce corpus, même si de fait, il y a bien prise en compte de 117900 contextes potentiellement tous différents. L'annotation retient pour les 24291 entrées non voyellées, 38108 descriptions morpho-grammaticales différentes, sur 334179 descriptions potentielles, c'est-à-dire qu'elle donne lieu à 38108 hyper-formes dûment voyellées, lemmatisées et étiquetées différentes.

7 Conclusion

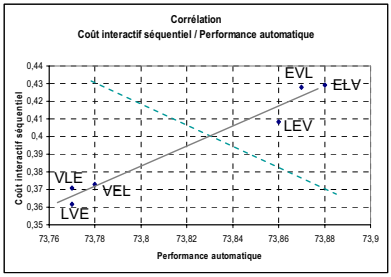


Figure 2.

Ce graphe, qui reprend les résultats lignes D et F de la figure 1, indique que, dans une plage de performances donnée, *le meilleur comportement automatique n'est pas celui qui assure toujours, dès lors qu'il y a intervention manuelle, le meilleur comportement interactif.* Dit autrement, et en soulignant le caractère local de nos observations, nous découvrons en effet que *le comportement autonome le plus performant n'est pas toujours celui qui garantit, dès lors qu'il y a interaction, le comportement coopératif le plus performant.* C'est ce résultat *empirique* qui ne laisse de surprendre – la corrélation *Coût-Performance* attendue était et reste en effet qu'à performance automatique meilleure corresponde coût interactif moindre, et que les points dessinent la courbe de tendance en pointillée, et non celle observée trait continu – qui est devenu, chemin faisant, prépondérant, au-delà de la conception et réalisation d'un système d'analyse morpho-grammaticale interactive de l'arabe et de sa mise en œuvre pour la confection de corpus voyellés, étiquetés, et lemmatisés. Sur le plan méthodologique, il remet en cause les stratégies communément préconisées pour la confection massive de données linguistiques annotées, où l'idée est de corriger les sorties d'analyse automatique au moyen d'éditeurs dédiés, en considérant *a priori* que le rendement optimal est atteint dès lors que l'analyseur automatique qui est mis en œuvre est le plus performant. Nous pressentons, sans l'avoir encore constaté, que cet *a priori* n'est vrai qu'au-delà d'un certain seuil de performance automatique, seuil qu'il conviendrait de déterminer. En deçà de ce seuil critique, nous assisterions à des comportements « erratiques » où précisément, ainsi que nous l'avons observé, la corrélation *meilleure performance automatique* alors *meilleure performance interactive* n'est pas maintenue. Au-delà, au contraire, la corrélation est ou serait rétablie.

Mais il nous semble que les conséquences de ce constat vont plus loin encore. S'il devait être confirmé par d'autres expérimentations menées par nous ou par d'autres, sur d'autres langues et/ou d'autres types de règles, alors nous serions fondés à dire que nos objectifs devraient non plus se focaliser sur la seule dimension automatique, ainsi que nous disions au début de notre introduction, mais devraient aussi, d'emblée, prendre en compte le développement de la

nécessaire dimension interactive et de ce que celle-ci induit dans le développement et l'évaluation de la dimension automatique. Car l'on s'aperçoit qu'introduire parallèlement une dimension interactive plus dynamique, loin de réduire la surface de la composante automatique ou de ce que l'on peut en exiger, conduit au contraire à en multiplier les comportements linguistiques et à en étendre les potentialités, tout en en révélant les insuffisances critiques. L'interactif ramène ainsi à l'automatique.

Sous l'angle de l'évaluation, l'interactif conduit, comme pour les applications qui mettent en œuvre différents composants linguistiques, et où l'on distingue, (cf. par ex. Berthelin et al. 2001), les performances intrinsèques de ces composants d'une part, et les performances globales de ces mêmes composants interagissant ensemble, à une caractérisation tierce de ces composants. Mais là s'arrête l'analogie. Les métriques restent en effet les mêmes dans le premier cas. Qu'il s'agisse de performances locales et directes, ou globales et indirectes, l'on essaie de compter les écarts, les erreurs, les silences. Alors qu'elles sont différentes lorsqu'il s'agit de mesurer les performances de la dimension interactive, ainsi que nous avons essayé de le montrer. L'on rejoint ici l'une des multiples « *dimensions* » de l'évaluation recensées par Chaudiron, en l'occurrence, la notion d'*efficience* vue, pour une tâche donnée, « *comme la possibilité pour un utilisateur d'accomplir cette tâche à moindre coût en terme de charge de travail et d'effort cognitif* » (Chaudiron, 2001, p. 100). Corrélée à l'évaluation de l'analyse automatique, l'évaluation de l'annotation interactive souligne au final qu'il est certes crucial de parfaire les performances de l'automatique, mais qu'il est aussi utile d'octroyer à celui-ci non plus un, mais différents comportements pour être à même de s'adapter de façon optimale à la variabilité des comportements des annotateurs, sous peine de ne pas être retenu.

Remerciements

Le présent travail a été initié dans le cadre du projet EurADic (Action Technolangue du Ministère de la recherche), et se poursuit dans le cadre du projet MUSCLE (6ème PCRD).
A J.-B. Berthelin, pour la traduction du résumé, et sa disponibilité à aborder ces thématiques.

Références

ABEILLE A., CLEMENT L. (2003). *Annotation morpho-syntaxique. Les mots simples – Les mots composés. Corpus Le Monde*. Technical report, Paris 7.

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(1).

BERTHELIN J.-B. (2001). Two levels of evaluation in a complex NL system. Actes d'*ACL'2001*, *Toulouse*.

CHAUDIRON S. (2001). *L'Évaluation des systèmes de traitement de l'information textuelle : vers un changement de paradigme*. Habilitation à diriger des recherches, Paris X, Nov. 2001.

DEBILI F., ACHOUR H., SOUISSI E. (2002). La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique. *Correspondances N°71, IRMC,* Tunis, 10-26.

DEBILI F., FLUHR C. (2006). Confection de ressources dictionnairiques et textuelles multilingues. Actes de *TALN'2006, Louvain, Belgique, 10-13 Avril 2006*, 910-917.

DEBILI F., SOUISSI E. (2005). Y a-t-il une taille optimale des règles de succession intervenant dans l'étiquetage grammatical ? Actes de *TALN'2005, Dourdan, Juin 2005*, 363-372.

HABERT B. (2005). *Instruments et ressources électroniques pour le français*. Paris : Editions Ophrys.

TESNIERE L. (1969). *Eléments de syntaxe structurale*. Paris : Editions Klincksieck.

VERONIS J. (1999). *Guide d'étiquetage Multitag*. Version 3.1, 6 novembre 1999.

Évaluation des stades de développement en français langue étrangère*

Jonas GRANFELDT[1], Pierre NUGUES[2]
[1] Centre de langues et de littérature, Université de Lund, S-221 00 Lund
[2] Institut d'informatique, Institut de Technologie de Lund, S-221 00 Lund
Jonas.Granfeldt@rom.lu.se, Pierre.Nugues@cs.lth.se

Résumé. Cet article décrit un système pour définir et évaluer les stades de développement en français langue étrangère. L'évaluation de tels stades correspond à l'identification de la fréquence de certains phénomènes lexicaux et grammaticaux dans la production des apprenants et comment ces fréquences changent en fonction du temps. Les problèmes à résoudre dans cette démarche sont triples : identifier les attributs les plus révélateurs, décider des points de séparation entre les stades et évaluer le degré d'efficacité des attributs et de la classification dans son ensemble. Le système traite ces trois problèmes. Il se compose d'un analyseur morphosyntaxique, appelé Direkt Profil, auquel nous avons relié un module d'apprentissage automatique. Dans cet article, nous décrivons les idées qui ont conduit au développement du système et son intérêt. Nous présentons ensuite le corpus que nous avons utilisé pour développer notre analyseur morphosyntaxique. Enfin, nous présentons les résultats sensiblement améliorés des classificateurs comparé aux travaux précédents (Granfeldt *et al.*, 2006). Nous présentons également une méthode de sélection de paramètres afin d'identifier les attributs grammaticaux les plus appropriés.

Abstract. This paper describes a system to define and evaluate stages of development in second language French. The task of identifying such stages can be formulated as identifying the frequency of some lexical and grammatical features in the learners' production and how they vary over time. The problems in this procedure are threefold : identify the relevant features, decide on cutoff points for the stages, and evaluate the degree of efficiency of the attributes and of the overall classification. The system addresses these three problems. It consists of a morphosyntactic analyzer called Direkt Profil and a machine-learning module connected to it. We first describe the usefulness and rationale behind the development of the system. We then present the corpus we used to develop our morphosyntactic analyzer called Direkt Profil. Finally, we present new and substantially improved results on training machine-learning classifiers compared to previous experiments (Granfeldt *et al.*, 2006). We also introduce a method of attribute selection in order to identify the most relevant grammatical features.

Mots-clés : analyseur morphosyntaxique, apprentissage automatique, acquisition des langues.

Keywords: morphosyntactic parser, machine learning, language acquisition.

* Une première version de cet article a été présentée à la 16e conférence nordique de traitement automatique des langues, NODALIDA 2007, Tartu, Estonie, sous le titre *Evaluating Stages of Development in Second Language French : A Machine-Learning Approach*.

1 Introduction

L'un des points essentiels des recherches sur l'acquisition des langues étrangères est l'identification et l'analyse des stades de développement que l'on traverse en apprenant une deuxième langue ou une langue étrangère. La notion de stade de développement peut s'appliquer aux données de tous les niveaux linguistiques, mais elle est particulièrement intéressante pour rendre compte de l'acquisition de la morphologie et de la syntaxe. Dans ce cadre, on peut considérer la grammaire interne de l'apprenant comme un système propre qui se développerait et subirait des restructurations au cours du temps.

La modélisation du développement de la grammaire de l'apprenant et de ses propriétés à des moments différents revient, pour l'essentiel, à identifier les phénomènes grammaticaux pertinents, à définir des points de séparation entre les stades et à les évaluer de manière systématique. Dans cet article, nous décrivons et nous évaluons un système qui a entièrement automatisé ce processus. Avant de le présenter, nous décrivons de façon simplifiée comment on identifie en général les stades de développement dans le domaine de l'acquisition des langues.

2 Contexte

2.1 Méthode actuelle pour identifier les stades de développement

La première étape pour identifier les stades de développement est de déterminer et d'extraire des phénomènes grammaticaux dans la production, orale ou écrite, d'une population représentative d'apprenants. Le point crucial dans le choix de ces phénomènes est qu'ils aient une validité interne, dont les réalisations peuvent traduire un changement qualitatif de la grammaire. Une deuxième étape est de comprendre et modéliser leur développement.

Certains phénomènes linguistiques montrent un développement linéaire simple et les pourcentages d'usages corrects ont une augmentation stable avec le temps. D'autres phénomènes ont un développement non linéaire, parfois en forme de « U », où les pourcentages d'usages corrects au début de l'acquisition sont élevés mais diminuent dans une deuxième phase pour ensuite regagner un niveau élevé de rectitude dans une troisième phase. Une explication qui a souvent été proposée pour ce type d'évolution est que la première phase contient un certain nombre d'expressions fixes apprises de façon holistique par l'apprenant. Ces structures, peut-être apprises par cœur, représenteraient alors des structures linguistiques non-analysées dans la grammaire interne des apprenants. Ensuite, une fois que les séquences de développement sont connues, il faut décider des points de séparation dans les données où l'apprenant a atteint un nouveau stade de développement. La plupart du temps, il est préférable de prendre en compte plusieurs phénomènes grammaticaux pour en même temps réaliser un « profilage grammatical ».

2.2 Problèmes de la méthode actuelle

L'analyse morphosyntaxique détaillée de textes ou d'énoncés produits par les apprenants est une partie centrale dans la méthode décrite dans le paragraphe précédent. La plupart des analystes travaillant sur l'acquisition d'une première et d'une deuxième langue ont maintenant accès à de grands corpus de productions orales et écrites. Dans notre cas, ce sont des textes écrits mais on

pourrait tout aussi bien l'appliquer à des transcriptions de productions orales. Pour des langues répandues, comme l'anglais et le français, on dispose également d'outils tels que des analyseurs morphologiques et des étiqueteurs de parties du discours (MacWhinney, 2000). Ces outils peuvent réduire de façon considérable le temps de l'analyse morphosyntaxique qui autrement serait très fastidieuse.

Cependant même avec ces outils, beaucoup d'analyses manuelles restent à faire. D'abord il n'existe actuellement aucun outil automatisé fiable pour l'analyse syntaxique de textes d'apprenants malgré quelques tentatives récentes pour l'anglais (Sagae *et al.*, 2005). Pour le français, une partie des structures linguistiques utilisées dans le profilage grammatical peuvent être détectées en utilisant des outils comme CHILDES. Mais pour d'autres structures plus complexes telles que l'accord entre les constituants, c'est impossible. Un autre problème est qu'avec les outils actuels, on ne peut effectuer que des requêtes simples sur un phénomène individuel alors que dans le profilage grammatical, on doit en analyser un grand nombre en même temps.

Un troisième problème concerne le côté artificiel des stades. Le résultat de l'analyse morphosyntaxique est présenté typiquement sous forme de fréquences de certains phénomènes. Pour un phénomène linguistique particulier, par exemple l'accord sujet-verbe à la troisième personne du singulier au présent, on identifie les différentes réalisations de cette structure et on les compte. Les données compilées pour tous les apprenants et tous les phénomènes et structures faisant partie du profil grammatical sont ensuite inspectées afin d'identifier intuitivement des stades de développement. Il y a actuellement de multiples façons de traiter cette étape, mais aucune n'a reçu d'évaluation systématique. Une raison possible est que personne n'ait relié les deux premières étapes, l'analyse morphosyntaxique et l'analyse des fréquences, à un traitement statistique. Si un traitement entièrement automatisé du processus était disponible, toutes ses étapes auraient pu être évaluées plus complètement.

Dans le reste de l'article, nous présentons notre système qui vise à surmonter les problèmes mentionnés précédemment. Nous commençons par un bref résumé des travaux précédents sur le développement morphosyntaxique du français langue étrangère. Nous décrivons ensuite le corpus que nous employons ainsi que de façon brève notre analyseur, Direkt Profil. Dans les derniers paragraphes, nous discutons de notre démarche fondée sur l'apprentissage automatique pour définir et évaluer les stades de développement et pour sélectionner les attributs. Nous présentons enfin nos résultats actuels.

3 Développement morphosyntaxique du français deuxième langue

Une partie des recherches sur le développement morphosyntaxique de français langue étrangère a pour objectif d'atteindre une description détaillée de la façon dont les apprenants développent leur grammaire en fonction du temps. L'étude de Bartning & Schlyter (2004) en est un exemple pour le français parlé, où les auteurs ont identifié environ 25 constructions morphosyntaxiques différentes et ont proposé une définition de leur développement dans le temps pour des Suédois adultes. Pris ensemble, ces phénomènes délimitent six stades sous la forme de profils grammaticaux qui s'étendent des débutants aux apprenants très avancés. Des exemples de constructions sont donnés dans le tableau 1. Au fur et à mesure que l'apprenant automatise la mise en œuvre de la langue cible, les structures produites deviennent plus fréquentes, plus complexes et plus correctes. Les itinéraires d'acquisition décrivent ce processus en termes linguistiques.

Stades	1	2	3	4	5	6
% Formes conjuguées de verbes lexicaux en contextes obligatoires	50-75	70-80	80-90	90-98	100	100
% Accord 1re personne pluriel S-V (*nous V-ons*)	–	70-80	80-95	100	100	100
% Accord 3e pers pluriel avec verbes irréguliers lexicaux comme *viennent*, *veulent*, *prennent*	–	–	qq cas	≈ 50	qq erreurs	100
Placement des pronoms objets	–	SVO	S(v)oV	SovV app.	SovV prod	acquis (*y* et *en*)
% Accord genre grammatical	55-75	60-80	65-85	70-90	75-95	90-100

TAB. 1 – Itinéraire de développement d'après Bartning & Schlyter (2004). Légende : – = pas d'occurrences ; app = apparaît ; prod = productif niveau avancé.

4 Un corpus écrit de français langue étrangère

Pour développer notre analyseur (voir § 5) et expérimenter notre approche d'apprentissage automatique des stades, nous avons utilisé le Corpus Écrit de Français Langue Étrangère de Lund (Ågren, 2005) – CEFLE. CEFLE se compose de textes en français provenant de 85 étudiants suédois à différents niveaux de développement. Il contient approximativement 400 textes et 100 000 mots. Il comporte également des textes d'un groupe de contrôle de 22 jeunes Français du même âge. CEFLE a été compilé lors de l'année scolaire 2003/2004 pendant laquelle chaque étudiant a écrit quatre ou cinq textes à deux mois d'intervalle.

Pour notre étude, nous avons utilisé un sous-ensemble de 317 textes du corpus dont les caractéristiques sont données dans le tableau 2. En employant les critères décrits dans Bartning & Schlyter (2004), un membre de l'équipe a au préalable annoté un texte de chaque étudiant et a estimé le stade de développement qu'il reflétait. Pour les expérimentations que nous décrivons dans les paragraphes qui suivent, nous avons attribué de façon systématique la même classification aux trois ou quatre autres textes du même étudiant dans le corpus. Nous avons ainsi propagé le stade annoté à la main à tous les textes du même étudiant. Nous avons supposé que généralement l'étudiant ne monterait pas d'un stade pendant la courte période où a eu lieu la collecte des textes.

CEFLE			**Sous-ensemble de CEFLE** (moyenne)			
Tâche	Type d'élicitation	Mots	Stade	Textes	Taille texte	Taille phrases
Homme	Images	17 260	Stade 1	23	78	6,9
Souvenir	Récit personnel	14 365	Stade 2	98	161	8,4
Italie	Images	30 840	Stade 3	97	212	9,8
Moi	Récit personnel	30 355	Stade 4	58	320	11,6
Total		92 820	Contrôle	41	308	15,2

TAB. 2 – La description générale du corpus CEFLE et du sous-ensemble utilisé dans les expériences rapportées dans cet article.

5 Direkt Profil

Direkt Profil (Granfeldt *et al.*, 2005, 2006) est un analyseur morphosyntaxique conçu pour du français langue étrangère. Le but initial était de mettre en œuvre une analyse automatique des phénomènes et des constructions grammaticaux contenus dans le tableau 1. Dans sa version actuelle, le système ne détecte pas certains des phénomènes indiqués par Bartning & Schlyter (2004), mais en contrepartie il en détecte un grand nombre d'autres. Le système a fait l'objet d'une présentation détaillée dans des articles précédents et nous nous contenterons d'en donner une brève description.

Le concept de groupe, nominal ou verbal, correct ou non, représente le support grammatical essentiel de notre analyse. Nous avons défini une annotation des textes, propre au projet, fondé sur ces groupes. Elle prend en compte les phénomènes linguistiques caractéristiques des itinéraires de développement d'après les catégories décrites par Bartning & Schlyter (2004). La version actuelle de Direkt Profil, V. 2.1, détecte trois types de groupes syntaxiques : nominaux non-récursifs, verbaux et prépositionnels.

L'analyseur utilise des règles écrites manuellement et s'appuie sur un lexique de formes fléchies. De façon conceptuelle, l'analyseur recherche des classes de structures syntagmatiques sans considérer leurs traits grammaticaux. Il identifie ensuite les structures progressivement en tentant d'affecter des valeurs à ces traits. La reconnaissance des limites des groupes se fait par un ensemble de mots vides et par des heuristiques à l'intérieur des règles. Direkt Profil applique en cascade trois ensembles de règles pour produire quatre niveaux d'annotations. Le premier ensemble segmente le texte en mots. Un ensemble intermédiaire identifie les expressions figées. Le troisième ensemble annote simultanément les parties du discours et les groupes. Finalement, le moteur crée un groupe de résultats relié au stade de l'apprenant. Il est à noter que le moteur n'annote pas tous les mots, ni tous les segments. Il ne considère que ceux qui sont pertinents pour la détermination du stade. Le moteur applique les règles de gauche à droite puis de droite à gauche pour résoudre certains problèmes d'accord.

La version actuelle de Direkt Profil est accessible en ligne à l'adresse www.rom.lu.se :8080/profil. La performance de la version 1.5.2 pour la détection des segments a été évaluée dans Granfeldt *et al.* (2005). Les résultats ont donné une moyenne harmonique F globale de précision et de rappel de 0,83.

6 Une méthode d'apprentissage automatique pour évaluer les stades de développement

À l'heure actuelle, il existe des quantités de méthodes pour définir les stades de développement, mais à notre connaissance aucune façon systématique pour les évaluer. Dans leur article, Bartning & Schlyter (2004) avaient défini six stades de développement. Le corpus CEFLE en utilise cinq attribués par un annotateur humain. Un problème essentiel dans cette dernière étape est que l'analyse montre une augmentation progressive des fréquences avec l'acquisition qui suggère plutôt un développement en continu que selon des stades discrets. D'une certaine manière, il faut donc accepter que n'importe quelle définition soit en partie arbitraire.

Dans notre système, l'analyse des fréquences des constructions grammaticales est obtenue automatiquement et correspond à la sortie de Direkt Profil. Elle forme le support qui permet d'établir

les stades de développement. Dans le paragraphe suivant, nous évaluons la probabilité de l'existence de ces cinq stades différents en utilisant des techniques d'apprentissage automatique.

6.1 Première expérience : classification utilisant tous les attributs

Comme conditions expérimentales, nous avons employé chacun des textes des 85 étudiants qui a reçu manuellement un stade de développement. Nous avons ensuite réutilisé la même classification pour les trois ou quatre autres textes du même étudiant dans le corpus, ce qui donne comme résultat 276 textes d'apprenant classifiés. Les 41 textes supplémentaires viennent du groupe de contrôle des natifs, ce qui aboutit à un total de 317 textes classifiés. La phase d'apprentissage induit automatiquement des classifieurs à partir des vecteurs de 142 attributs que nous extrayons des textes au moyen de l'analyseur.

Nous avons employé trois algorithmes d'apprentissage automatique : ID3/C4.5 (Quinlan, 1986), les machines à vecteurs de support (SVM) (Boser *et al.*, 1992) et les arbres de modèles logistiques (LMT) (Landwehr *et al.*, 2003). Dans un premier temps, nous avons regroupé les cinq stades dans trois stades plus généraux, où les stades 1 et 2 ainsi que les stades 3 et 4 ont été fusionnés et nous avons entraîné les algorithmes sur ces trois stades. Nous avons ensuite réalisé une deuxième évaluation avec les cinq stades d'origine. Nous avons réalisé toutes nos expériences avec l'ensemble d'algorithmes d'apprentissage automatique disponible dans Weka[1] (Witten & Frank, 2005) et nous les avons évalués en appliquant 10 fois une validation croisée sur le corpus d'apprentissage. Les tableaux 3 et 4 présentent les résultats pour les 317 textes pour 3 et 5 classes respectivement.

	C4.5			SVM			LMT		
Stades	P	R	F	P	R	F	P	R	F
1–2	0,66	0,70	0,68	0,70	0,71	0,71	0,76	0,75	**0,75**
3–4	0,70	0,68	0,69	0,71	0,72	0,71	0,76	0,79	**0,77**
Contrôle	0,71	0,66	0,68	0,70	0,63	0,67	0,89	0,83	**0,86**

TAB. 3 – Résultats de la classification des textes en trois stades pour les trois classifieurs. Chaque classifieur a employé 142 attributs et a été entraîné sur 317 textes du corpus CEFLE. P : Précision. R : Rappel, F : Moyenne harmonique de la précision et du rappel.

	C4.5			SVM			LMT		
Stades	P	R	F	P	R	F	P	R	F
1	0,37	0,42	0,39	0,54	0,58	**0,56**	0,44	0,33	0,38
2	0,50	0,52	0,51	0,60	0,60	**0,60**	0,59	0,61	**0,60**
3	0,42	0,46	0,44	0,45	0,46	0,45	0,51	0,54	**0,53**
4	0,48	0,38	0,42	0,52	0,50	0,51	0,64	0,66	**0,65**
Contrôle	0,71	0,66	0,68	0,70	0,63	0,67	0,89	0,83	**0,86**

TAB. 4 – Résultats de la classification des textes en cinq stades pour les trois classifieurs. Chaque classifieur a employé 142 attributs et a été entraîné sur 317 textes du corpus CEFLE.

[1] Accessible à partir de ce site : http ://www.cs.waikato.ac.nz/ml/weka/.

Ces résultats peuvent être comparés à ceux que nous avons obtenus avec une version précédente de Direkt Profil (1.5.4) employant un nombre plus petit de 33 attributs et un corpus d'entraînement moins important de 80 textes. Ces résultats (Granfeldt *et al.*, 2006) ont montré que le meilleur classifieur, SVM, obtenait une moyenne harmonique de précision et de rappel de près de 70% pour la classification en trois stades, et une moyenne de 43% de précision et 36% de rappel pour une classification en cinq stades. Les résultats actuels avec plus de 100 attributs supplémentaires et un corpus d'entraînement qui est quatre fois plus grand montrent une amélioration de presque 10%. Le meilleur algorithme, cette fois LMT, obtient une moyenne de précision et de rappel de 79% pour la classification en trois stades (Tableau 3). Pour la classification en cinq stades, l'amélioration est encore plus importante (Tableau 4). LMT obtient 62% de précision et 59% de rappel. En comparant la performance des deux meilleurs algorithmes, SVM et LMT, nous observons que LMT est supérieur à SVM sur les stades intermédiaire et avancé – 3, 4, et le groupe de contrôle des natifs – mais pas sur les deux premiers stades de développement. Nous n'avons aucune explication pour ce fait.

En conclusion de cette première expérience, nous pouvons affirmer que le plus grand nombre d'attributs et le corpus d'entraînement plus important ont eu comme résultat une meilleure performance globale pour les trois classifieurs. Mais l'amélioration n'a pas été aussi grande qu'espérée. Une hypothèse possible est que nous avons introduit un certain nombre d'attributs non pertinents parmi les quelques 100 nouveaux. Pour cette raison, nous avons appliqué une procédure de sélection afin d'identifier les meilleurs attributs. Les résultats de cette deuxième expérience sont présentés dans le paragraphe suivant.

6.2 Deuxième expérience : classification utilisant une sélection d'attributs

Pour évaluer les 142 attributs, nous avons mesuré le gain d'information (Quinlan, 1986) pour chaque attribut par rapport à la classe. Ce critère est à la base de l'algorithme ID3 et fait partie de la boîte à outils Weka. Nous avons employé la méthode de recherche de rang qui classe les différents attributs en fonction de leur évaluation. Les tableaux 5 et 6 présentent les résultats pour respectivement les 10 et 20 meilleurs attributs selon cette méthode. Dans un deuxième temps, nous avons réalisé deux nouvelles classifications en utilisant les mêmes algorithmes que dans la première expérience et le même choix de 317 textes du corpus, mais cette fois avec un nombre d'attributs réduit aux meilleurs d'entre eux.

Lors de la première classification, nous avons évalué la performance des classifieurs en utilisant les 10 meilleurs attributs. Les résultats produits sont mitigés (voir le tableau 7 pour la classification en cinq stades). En moyenne, la réduction radicale du nombre d'attributs de 142 à 10 ne semble pas beaucoup affecter les résultats. Les moyennes des précision et rappel pour LMT sont respectivement de 66% et 58%. Ceci suggère qu'il y a beaucoup de bruit dans les 132 attributs restants. En revanche, les résultats pour le premier stade de développement se détériorent de façon importante. L'algorithme SVM n'identifie plus un seul texte au stade 1. Ceci suggère que le reste des 132 attributs contient des informations très importantes pour identifier ce stade. Dans notre deuxième évaluation, nous avons incorporé les 10 attributs suivants (attributs 11–20, soit au total 20 attributs). Les résultats pour la classification en cinq stades sont présentés dans le tableau 8.

Les résultats globaux de cette deuxième classification sont meilleurs et le vecteur de 20 attributs permet à chacun des trois classifieurs d'identifier des textes au stade 1. Cependant la moyenne pour LMT est en légère baisse par rapport à la classification avec 10 attributs. On note également

Mérite	Rang	Attribut
0,405	1,4	Pourcentage de séquences déterminant-nom avec accord (nombre et genre)
0,354	2,2	Pourcentage de mots inconnus
0,33	3,2	Pourcentage de GNs avec accord en genre
0,313	3,9	Pourcentage de prépositions (sur toutes les parties de discours)
0,311	4,3	Longueur moyenne des phrases
0,208	6,2	Pourcentage de séquences nom-adjectif avec accord (nombre et genre)
0,198	7,4	Pourcentage d'accord sujet-verbe avec des verbes modaux + INF
0,187	8,3	Pourcentage d'accord sujet-verbe avec verbes au passé composé
0,177	9,3	Pourcentage d'accord sujet-verbe avec être/avoir au 3e personne pluriel
0,176	9,8	Pourcentage d'accord sujet-verbe avec verbes modaux et sujets pronominaux

TAB. 5 – Les 10 meilleurs attributs. Attributs 1–10.

Mérite	Rang	Attribut
0,168	11,4	Pourcentage de verbes au présent (sur tous les temps)
0,165	11,8	Pourcentage de verbes au passé composé (sur tous les temps)
0,15	14	Pourcentage d'accord sujet-verbe avec aux. de mode (sur tous les sujets)
0,142	15,7	Pourcentage d'accord sujet-verbe avec aux. de mode au sg.
0,14	16,2	Pourcentage d'accord sujet-verbe avec aux. de mode au présent et sujet pronominal 3e personne
0,136	16,7	Pourcentage verbes lexicaux conjugués dans des contextes conjugués
0,133	17,3	Pourcentage d'accord sujet-verbe avec verbes lexicaux conjugués
0,131	18,1	Pourcentage d'accord sujet-verbe avec sujet pronominal sg et aux. de mode
0,125	19,3	Pourcentage d'accord sujet-verbe avec verbes lexicaux à la 3e personne du pluriel
0,116	21,4	Pourcentage d'accord sujet-verbe avec sujet pronominal et être/avoir

TAB. 6 – Les 10 attributs suivants. Attributs 11–20.

	C4.5			SVM			LMT		
Stades	P	R	*F*	P	R	*F*	P	R	*F*
1	0,46	0,46	**0,46**	0,00	0,00	0,00	0,78	0,29	0,42
2	0,50	0,49	0,49	0,53	0,72	0,61	0,57	0,70	**0,63**
3	0,43	0,42	0,43	0,50	0,43	0,46	0,55	0,49	**0,52**
4	0,50	0,57	0,53	0,62	0,71	**0,66**	0,63	0,64	0,63
Contrôle	0,84	0,76	0,79	0,94	0,76	**0,84**	0,78	0,78	0,78

TAB. 7 – Résultats de la classification des textes en cinq stades pour les trois classificateurs. Chaque classificateur a employé les 10 meilleurs attributs évalués avec le critère du gain d'information de Weka et a été entraîné sur 317 textes du corpus.

	C4.5			SVM			LMT		
Stades	P	R	*F*	P	R	*F*	P	R	*F*
1	0,56	0,38	0,45	0,60	0,38	**0,46**	0,53	0,38	0,44
2	0,51	0,53	0,52	0,61	0,62	**0,62**	0,61	0,61	0,61
3	0,49	0,47	0,48	0,54	0,57	0,56	0,56	0,59	**0,57**
4	0,45	0,55	0,50	0,61	0,69	**0,65**	0,61	0,62	0,62
Contrôle	0,78	0,68	0,73	0,83	0,73	0,78	0,86	0,88	**0,87**

TAB. 8 – Résultats de la classification des textes en cinq stades pour les trois classificateurs. Chaque classificateur a employé les 20 meilleurs attributs évalués avec le critère du gain d'information de Weka et a été entraîné sur 317 textes du corpus.

une différence dans les chiffres de précision et de rappel pour le stade 1, cette fois-ci par rapport à la première expérience utilisant les 142 attributs (voir les tableaux 3 et 4). Alors que ces chiffres étaient relativement proches, ils sont très différents dans les deux expériences suivantes avec un rappel considérablement inférieur à la précision. Ceci signifie que la qualité du rappel dépend d'un ensemble d'attributs beaucoup plus grand pour le stade de développement le plus bas que pour les autres stades. Puisque la précision et le rappel pour les stades plus avancés sont proches dans toutes expériences menées, ceci pourrait signifier que le stade 1 est le plus hétérogène.

7 Conclusion

Dans cet article, nous avons présenté et évalué un système pour identifier des stades de développement en français langue étrangère. Le système se compose d'un analyseur morphosyntaxique et d'un module d'apprentissage automatique. Dans un premier temps, l'analyse morphosyntaxique nous permet de représenter chaque texte par un vecteur de 142 attributs. Grâce aux second module, nous avons ensuite entraîné trois classifieurs différents pour évaluer l'hypothèse qu'on pouvait partitionner les textes du corpus en cinq stades de développement. Cette démarche a permis la classification automatique d'un ensemble de 317 textes du corpus CEFLE selon le stade de développement qu'ils reflétaient.

Les résultats d'une première expérience de classification employant un vecteur contenant l'ensemble des 142 attributs ont montré une amélioration importante de plus de 10% comparés à nos résultats précédents. Pour une classification simplifiée à trois stades, la moyenne de précision et de rappel pour le système est maintenant de 79%. Dans le but d'identifier les meilleurs attributs pour la classification, nous avons introduit un critère de sélection fondée sur le gain d'information. À notre surprise, les résultats ont montré que la performance globale n'était pas affectée de façon sensible par la réduction radicale du nombre d'attributs (de 142 à 10 et 20 respectivement). Cependant les résultats pour le stade de développement le plus bas sont dégradés de façon très importante. Une interprétation possible est que les textes du stade 1 sont tellement hétérogènes qu'on doive remettre en cause l'unicité de ce niveau et qu'il serait préférable de le diviser en plusieurs sous-classes.

Remerciements

La recherche présentée dans cet article bénéficie d'un financement du Conseil suédois pour la science, contrat numéro 2004-1674, et de bourses de la fondation Elisabeth Rausing pour la recherche dans les sciences humaines et de la fondation Erik Philip-Sörenssen pour la recherche.

Références

ÅGREN M. (2005). *Le marquage morphologique du nombre dans la phrase nominale. Une étude sur l'acquisition du français L2 écrit*. Rapport interne, Institut d'études romanes de Lund. Université de Lund.

BARTNING I. & SCHLYTER S. (2004). Stades et itinéraires acquisitionnels des apprenants suédophones en français L2. *Journal of French Language Studies*, **14**(3), 281–299.

BOSER B., GUYON I. & VAPNIK V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, p. 144–152, Pittsburgh : ACM.

GRANFELDT J., NUGUES P., PERSSON E., PERSSON L., KOSTADINOV F., ÅGREN M. & SCHLYTER S. (2005). Direkt Profil : un système d'évaluation de textes d'élèves de français langue étrangère fondé sur les itinéraires d'acquisition. In *Actes de la conférence Traitement Automatique des Langues Naturelles, TALN & RECITAL 2005*, volume Tome 1 – Conférences principales, p. 113–122, Dourdan, France.

GRANFELDT J., NUGUES P., ÅGREN M., THULIN J., PERSSON E. & SCHLYTER S. (2006). CEFLE and Direkt Profil : A new computer learner corpus in French L2 and a system for grammatical profiling. In *Proceedings of the 5th International Conference on Language Ressources and Evaluation*, p. 565–570, Genoa, Italy.

LANDWEHR N., HALL M. & FRANK E. (2003). Logistic model trees. In N. LAVRAC, D. GAMBERGER, L. TODOROVSKI & H. BLOCKEEL, Eds., *Proceedings of the 14th European Conference on Machine Learning (ECML)*, volume 2837 of *Lecture Notes in Computer Science*, p. 241–252. Springer.

MACWHINNEY B. (2000). *The CHILDES project : Tools for analyzing talk*. Mahwah, New Jersey : Lawrence Erlbaum.

QUINLAN J. R. (1986). Induction of decision trees. *Machine Learning*, **1**(1), 81–106.

SAGAE K., LAVIE A. & MACWHINNEY B. (2005). Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics 2005*, p. 197–2004, Ann Arbor, USA.

WITTEN I. H. & FRANK E. (2005). *Data Mining. Practical Machine Learning Tools and Techniques*. Amsterdam : Elsevier.

Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique

Delphine Bernhard
TIMC-IMAG, Institut d'Ingénierie de l'Information de Santé
Faculté de Médecine, 38706 La Tronche cedex
Delphine.Bernhard@imag.fr

Résumé. Cet article présente un système d'acquisition de familles morphologiques qui procède par apprentissage non supervisé à partir de listes de mots extraites de corpus de textes. L'approche consiste à former des familles par groupements successifs, similairement aux méthodes de classification ascendante hiérarchique. Les critères de regroupement reposent sur la similarité graphique des mots ainsi que sur des listes de préfixes et de paires de suffixes acquises automatiquement à partir des corpus traités. Les résultats obtenus pour des corpus de textes de spécialité en français et en anglais sont évalués à l'aide de la base CELEX et de listes de référence construites manuellement. L'évaluation démontre les bonnes performances du système, indépendamment de la langue, et ce malgré la technicité et la complexité morphologique du vocabulaire traité.

Abstract. This article describes a method for the unsupervised acquisition of morphological families using lists of words extracted from text corpora. It proceeds by incrementally grouping words in families, similarly to agglomerative hierarchical clustering methods. Clustering criteria rely on graphical similarity as well as lists of prefixes and suffix pairs which are automatically acquired from the target corpus. Results obtained for specialised text corpora in French and English are evaluated using the CELEX database and manually built reference lists. The evaluation shows that the system perfoms well for both languages, despite the morphological complexity of the technical vocabulary used for the evaluation.

Mots-clés : familles morphologiques, classification, apprentissage non supervisé.

Keywords: morphological families, clustering, unsupervised learning.

1 Introduction

L'analyse morphologique est une tâche importante dans divers domaines du traitement automatique des langues comme la reconnaissance de la parole, la communication alternative et augmentée, la traduction automatique ou la recherche d'informations. Dans ce dernier cas, l'utilité des connaissances morphologiques se justifie par la proximité sémantique des variantes flexionnelles ou dérivationnelles. Il est également possible d'exploiter

la structure morphologique des mots pour l'acquisition de relations sémantiques telles que l'hyponymie (Buitelaar & Sacaleanu, 2002) ou l'antonymie (Schwab *et al.*, 2005). Les ressources décrivant les liens morphologiques n'étant pas disponibles à l'heure actuelle pour toutes les langues et tous les domaines, ces applications sont fréquemment associées à l'acquisition automatique de connaissances morphologiques à partir de textes. Les méthodes d'analyse morphologique non supervisée sont variées : comparaison de graphies (Zweigenbaum & Grabar, 2000), recherche d'analogies (Lepage, 1998), modèles probabilistes (Creutz & Lagus, 2005) ou segmentation par optimisation (Goldsmith, 2001; Creutz & Lagus, 2002). Elles se distinguent également par le type de résultats obtenus : mots découpés en segments morphémiques ou liens morphologiques.

Le travail présenté dans cet article relève du second type de méthode car il consiste en l'acquisition de familles morphologiques, c'est-à-dire des groupes de mots liés deux à deux par un lien morphologique d'affixation (préfixation ou suffixation) ou de composition. Nous formulons la question de l'acquisition de familles morphologiques comme un problème de classification. En effet, l'objectif de la classification est d'organiser un ensemble de données en groupes homogènes et contrastés : dans notre cas, les groupes souhaités sont des familles de mots morphologiquement reliés. La méthode que nous proposons prend pour point de départ une liste de mots et les groupe en familles d'une manière similaire aux méthodes de classification ascendante hiérarchique utilisées en analyse de données. Elle a de plus la particularité d'être non supervisée et n'est donc pas liée à une langue ou à un domaine précis.

Nous allons dans un premier temps décrire les diverses étapes de la méthode avant de présenter et d'analyser les résultats obtenus pour des corpus de textes techniques (médecine et volcanologie) en français et en anglais. Nous nous intéressons plus particulièrement au vocabulaire technique car il se caractérise par l'utilisation fréquente des procédés de composition et de dérivation, notamment par préfixation.

2 Description de la méthode

Le système prend pour entrée les données suivantes :

- Une liste des mots d'un corpus L
- Une liste de préfixes P
- Une liste de signatures (ou paires de suffixes) S

Les deux dernières listes sont obtenues à partir de la première à l'aide du module d'apprentissage d'affixes décrit dans (Bernhard, 2006). Celui-ci utilise les probabilités transitionnelles entre sous-chaînes pour repérer les zones de faible probabilité et ainsi découper les mots en radical et affixes. Nous avons adapté ce module pour qu'il produise non seulement une liste de préfixes et de suffixes mais également une liste de paires de suffixes qui apparaissent avec la même base et qui sont donc mutuellement substituables sur l'axe paradigmatique[1]. Par exemple, les suffixes de la paire (*s*,*ique*) peuvent se combiner à la base *climat* pour former les mots *climat**s*** et *climat**ique***. La même signature se retrouve dans les paires de mots *volcan**s*** – *volcan**ique*** et *océan**s*** – *océan**ique***. La notion de signa-

[1]Il faut noter que les préfixes et les suffixes sont acquis automatiquement, de manière non supervisée. Par conséquent, aucune distinction n'est faite entre les affixes flexionnels et dérivationnels.

ture est présente dans de nombreux travaux en acquisition automatique de connaissances morphologiques, parfois sous des dénominations différentes : *paires de suffixes* (Gaussier, 1999), *règles morphologiques* (Grabar & Zweigenbaum, 1999) ou *schémas de suffixation* (Hathout, 2005).

Nous allons maintenant détailler l'ensemble des étapes menant à l'acquisition des familles morphologiques.

2.1 Familles initiales

Avant apprentissage, il y a autant de familles que de mots dans la liste donnée en entrée : chaque mot constitue sa propre famille. Les familles formées au cours du processus d'apprentissage sont représentées par un radical R. De plus, chaque famille comprend deux sous-familles, sauf si elle correspond à une feuille dans la hiérarchie : dans ce cas, elle contient un mot unique et n'a pas de sous-famille.

2.2 Étape 1 : regroupement de familles à partir de l'inclusion de mots

Le premier critère de regroupement des familles est l'inclusion de mots : il s'agit de repérer les mots formés par préfixation à partir d'un autre mot de la liste, selon une procédure détaillée ci-dessous :

Soient :
– $m_1, m_2, \ldots, m_i$ et m_j des mots de longueur minimale égale à 4 ;
– $F_1, F_2, \ldots, F_i$ des familles telles que $F_1 = [m_1], F_2 = [m_2], \ldots, F_i = [m_i]$;
– F_j une famille telle que $F_j = [m_j]$.
Les familles $F_1, F_2, \ldots, F_i$ et F_j sont regroupées pour former une nouvelle famille F_k si $m_1 = E_1 + m_j, m_2 = E_2 + m_j, \ldots, m_i = E_i + m_j$
où E_1, E_2, …, E_i représentent une suite maximale d'un ou plusieurs préfixes de la liste P, éventuellement séparés par des tirets, tels que chaque préfixe ait une longueur minimale de 3.
Le radical de la nouvelle famille F_k est m_j.

Par exemple, si F_1 = [sub-océaniques], F_2 = [océaniques] et F_3 = [intra-océaniques] alors il est possible de former une nouvelle famille F_4 telle que $F_4 = F_1 \cup F_2 \cup F_3$ = [sub-océaniques, océaniques, intra-océaniques]. En effet, les mots *sub-océaniques* et *intra-océaniques* contiennent tous le mot *océaniques*. De plus, ils débutent par les préfixes *sub+* et *intra+*. Le radical de la nouvelle famille est *océaniques*.

2.3 Étape 2 : regroupement de familles à partir des préfixes

Après avoir procédé à un premier regroupement des mots en fonction des mots inclus, nous utilisons d'autres critères de regroupement, basés sur la comparaison des graphies des radicaux des familles existantes et des préfixes auxquels ils peuvent être associés. En

effet, lorsque deux mots partagent un même préfixe et que leurs bases sont graphiquement similaires, alors il y a de fortes chances pour qu'ils soient également morphologiquement liés. Prenons l'exemple des mots suivants : *neuro-oncologist* et *neuro-oncology*. Ces deux mots débutent tous deux par le préfixe *neuro-* suivi d'une même chaîne de caractères de longueur 7 : *oncolog*. La combinaison de deux indices, à savoir le partage d'un préfixe, suivi d'une chaîne commune, est un indice suffisant dans la plupart des cas pour conclure que les mots sont morphologiquement liés.

Nous appliquons ces remarques de la manière suivante :

Soient :
- F_1 et F_2 deux familles ;
- R_1 le radical représentant F_1 ;
- R_2 le radical représentant F_2.

Les deux familles F_1 et F_2 sont regroupées dans une nouvelle famille F_3 ssi :

1. $R_1 = \alpha + s_1$ et $R_2 = \alpha + s_2$, où α est une chaîne de caractères de longueur minimale égale à 4 et s_1 et s_2 sont des chaînes de caractères différant au moins par leur premier caractère.
2. Il existe au moins un mot $m_1 \in F_1$ et un mot $m_2 \in F_2$ tels que m_1 et m_2 incluent le même préfixe.

Le radical R_3 de la nouvelle famille F_3 est le mot le plus court parmi R_1 et R_2.

Par exemple, si :

- F_1 = [océanique, intra-océanique] avec R_1 = *océanique* ;
- F_2 = [océaniques, sub-océaniques, intra-océaniques] avec R_2 = *océaniques*

alors il est possible de former une nouvelle famille :

$F_3 = F_1 \cup F_2$ = [océanique, intra-océanique, océaniques, sub-océaniques, intra-océaniques].

En effet, R_1 et R_2 partagent une chaîne initiale commune de longueur 9, *océanique*, et les mots *intra-océanique* de F_1 et *intra-océaniques* de F_2 ont en commun le préfixe *intra*. Le radical de F_3 est le radical le plus court, à savoir *océanique*.

2.4 Étape 3 : regroupement de familles à partir des signatures

La dernière étape de la classification consiste à utiliser la liste de signatures S donnée en entrée et à découvrir de nouvelles signatures à partir des regroupements opérés lors des étapes précédentes. Ces signatures vont permettre à leur tour d'effectuer de nouveaux regroupements, selon le principe du bootstrapping. Le processus se termine lorsqu'il n'est plus possible de découvrir de nouvelles signatures.

2.4.1 Découverte de nouvelles signatures

La découverte de nouvelles signatures se fait à partir des familles déjà constituées au cours des étapes précédentes. Les mots non préfixés de chaque famille sont comparés deux à deux afin d'obtenir une liste de signatures, selon la méthode suivante :

Soient m_1 et m_2 deux mots non préfixés appartenant à la famille F
tels que $m_1 = \alpha + s_1$ et $m_2 = \alpha + s_2$
avec $|\alpha| \geq 4$ et s_1 et s_2 des chaînes de caractères différant au moins par leur premier caractère.
Nous appellerons signature la paire de suffixes (s_1,s_2) et $sig(F, F)$ l'ensemble des signatures formées à partir d'une famille F, c'est-à-dire par comparaison bijective des mots non préfixés de F. Toutes ces signatures sont ajoutées à la liste des signatures S.

Prenons l'exemple de la famille suivante, formée lors des étapes 1 et 2 :

[trachyandésite, andésite, trachy-andésite, andésites, trachy-andésites, trachyandésites, andésitique, trachy-andésitique, trachyandésitique, trachy-andésitiques, trachyandésitiques, andésitiques].

La comparaison des graphies des mots non préfixés de cette famille conduit à l'identification des paires de suffixes suivantes : (ϵ, s), $(e, ique)$, $(e, iques)$, $(es, ique)$ et $(es, iques)$ (voir Figure 1).

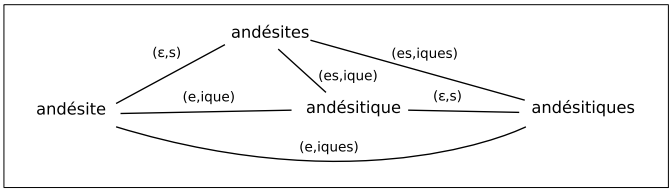


FIG. 1: Identification de signatures

2.4.2 Fusion de familles à l'aide des signatures

Les signatures ainsi acquises sont utilisées pour fusionner des familles. Le critère d'agglomération repose sur un indice p qui mesure la proportion de signatures valides partagées entre deux familles que l'on cherche à fusionner :

Soient :
- F_1 et F_2 deux familles ;
- l_1 le nombre de mots non préfixés de F_1 ;
- l_2 le nombre de mots non préfixés de F_2 ;
- S la liste de signatures fournies en entrée et découvertes à partir des familles déjà constituées.

$$p = \frac{|sig(F_1, F_2) \cap S|}{l_1 \cdot l_2}$$

Dans les expériences relatées dans la suite de cet article, nous avons fusionné deux familles lorsque $p \geq 0.5$.

Prenons l'exemple des familles représentées sur la Figure 2. Les signatures connues sont représentées par un arc plein tandis que les signatures inconnues sont représentées en pointillés. Ces deux familles sont fusionnées car le rapport du nombre de signatures connues sur le nombre total de signatures possibles est égal à 0.5.

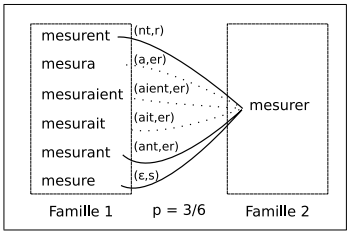


Fig. 2: Fusion de familles

Le dendrogramme de la Figure 3 illustre l'intérêt des regroupements effectués aux diverses étapes de la méthode. La seule famille formée à l'issue des deux premières étapes est [satellites, microsatellites, microsatellite, sous-satellite, satellite, mini-satellite]. L'étape de fusion de familles à partir des signatures partagées permet le regroupement de mots comme [satellitaire, satellitaires] ou [satellisation, satellisait].

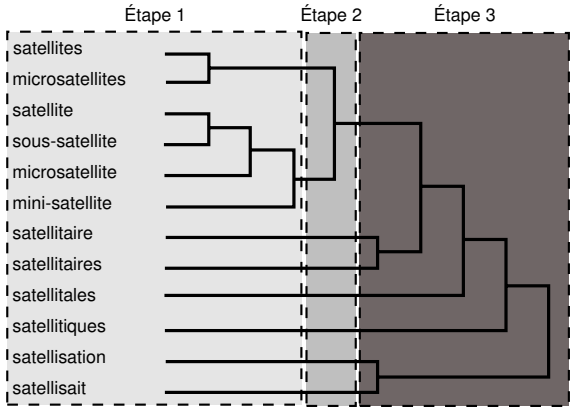


Fig. 3: Familles obtenues par classification à l'issue des trois étapes.

L'étape 3 d'agglomération à partir des signatures partagées est répétée tant que de nouvelles signatures sont acquises à partir des regroupements effectués et tant que ces signatures permettent de regrouper des familles. Le nombre de signatures différentes augmente fortement au cours des premières itérations, puis se stabilise. Le processus d'acquisition de nouvelles signatures, et par conséquent d'apprentissage de familles, s'achève au bout de 10 à 15 itérations.

3 Évaluation

Afin d'évaluer les résultats de la méthode, nous avons utilisé 4 corpus différents, en anglais et en français, couvrant deux domaines spécialisés distincts, la volcanologie et le cancer du sein. Dans la suite de cet article, ils seront désignés respectivement par `volcano-en`, `volcano-fr`, `cancer-en` et `cancer-fr`. Ces corpus ont été construits automatiquement à partir du Web en utilisant la méthode décrite dans (Baroni & Bernardini, 2004). Les listes de mots extraites de ces corpus comprennent entre 47 000 et 86 000 formes différentes.

3.1 Méthode d'évaluation

L'évaluation des résultats nécessite de disposer de familles morphologiques de référence auxquelles sont comparées les familles obtenues automatiquement par classification. Nous avons utilisé deux sources pour les familles de référence : nous avons d'une part élaboré manuellement des listes de référence et, pour l'anglais, nous avons extrait des familles de référence à partir des segmentations contenues dans la base CELEX (Baayen *et al.*, 1995). Les listes de référence construites manuellement contiennent des familles de mots pour le domaine du cancer du sein en français et en anglais. Elles contiennent 3 250 familles en anglais et 1 964 familles en français. Les familles morphologiques maximales de CELEX sont déterminées à partir des relations morphologiques de dérivation, de composition et de conversion, ce qui permet d'obtenir 14 880 familles de référence.

Nous avons évalué les familles induites par rapport aux familles de référence en utilisant les mesures proposées par (Schone & Jurafsky, 2000; Schone & Jurafsky, 2001). La méthode d'évaluation consiste à faire la somme des proportions de mots corrects (C), insérés (I) et supprimés (D) dans les familles morphologiques de tous les mots w de la liste d'évaluation. Si X_w est l'ensemble des mots appartenant à la famille morphologique d'un mot w selon le système à évaluer et Y_w est l'ensemble des mots appartenant à la famille morphologique de w selon CELEX ou toute autre base de référence, alors :

$$C = \sum_{\forall w} \frac{|X_w \cap Y_w|}{|Y_w|} \quad ; \quad D = \sum_{\forall w} \frac{|Y_w - (X_w \cap Y_w)|}{|Y_w|} \quad et \quad I = \sum_{\forall w} \frac{|X_w - (X_w \cap Y_w)|}{|Y_w|}$$

À partir de ces valeurs, il est également possible de calculer la précision, le rappel (et par conséquent la F-mesure) du système. La précision est égale à $C/(C+I)$ et le rappel à $C/(C+D)$.

3.2 Résultats obtenus

Les résultats sont détaillés dans la Table 1[2]. Ils démontrent la grande précision du système, qui est d'environ 80-90% suivant le corpus. De plus, malgré les contraintes imposées sur la longueur minimale des préfixes et des bases, le rappel est assez élevé et se situe autour

[2]Ces résultats ont été obtenus pour une valeur du paramètre N du module d'apprentissage des affixes égale à 10. N est un paramètre permettant de contrôler le processus d'apprentissage des affixes. Plus N est grand, plus le nombre de préfixes, de suffixes et, par conséquent, de signatures, est important.

de la barre des 70%. Ce résultat est d'autant plus remarquable que la méthode ne traite pas le cas des mots composés. La F-mesure varie peu sur l'ensemble des corpus, ce qui montre que les principes de regroupement utilisés sont valables aussi bien pour l'anglais que pour le français, et pour des domaines différents.

Référence	CELEX		Listes construites manuellement			
Corpus	cancer-en	volcano-en	cancer-en	volcano-en	cancer-fr	volcano-fr
Précision	79.3	81.4	89.8	91.2	91.5	93.1
Rappel	72.9	73.2	71.4	75.0	69.3	73.3
F-mesure	75.9	77.1	79.5	82.3	78.9	82.0

Tab. 1: Résultats obtenus pour les différents corpus et familles de référence.

3.3 Analyse des résultats

L'examen plus appronfondi des résultats montre que différents types de variantes sont groupés par l'algorithme :
- variantes orthographiques comme *tumor* (variante américaine) et *tumour* (variante britannique).
- variantes flexionnelles comme *traitement* et *traitements*.
- variantes dérivationnelles suffixées comme *traiter* et *traitement* et préfixées comme *auto-examen* et *examen*.
- composés savants comme *hormonothérapie* et *immunothérapie*. Il faut toutefois noter que dans ce cas, les chaînes *hormono* et *immuno* sont considérées comme des préfixes, car la méthode ne traite pas explicitement de la composition.

Malgré sa bonne précision, le système commet deux types d'erreur : sur-regroupement, c'est-à-dire le groupement de mots qui n'appartiennent pas tous à la même famille morphologique et sous-regroupement, c'est-à-dire l'absence de regroupement pour des mots appartenant à la même famille. La première erreur a pour conséquence de faire baisser la précision du système, tandis que la seconde conduit à une baisse du rappel. Ces erreurs peuvent survenir à toutes les étapes de la classification :

- À l'étape 1, malgré les contraintes imposées sur la longueur des préfixes et des mots, des mots peuvent être injustement considérés comme étant formés par combinaison d'un préfixe et d'un autre mot. Ainsi, le mot anglais *missing* est analysé comme étant la forme préfixée par *mis* du mot *sing*.
- À l'étape 2, il arrive que des familles soient fusionnées alors même qu'elles sont morphologiquement disjointes. Par exemple, la famille [médiane, paramédiane] est fusionnée avec la famille [socio-médical, paramédical, médical] car les mots *médiane* et *médical* commencent par la même chaîne de caractères de longueur 4 et apparaissent tous deux sous forme préfixée avec *para*.
- À l'étape 3 enfin, les contraintes imposées sur la longueur minimale de la base, égale à 4, peuvent empêcher le regroupement de certaines familles comme [île] et [îles], et donc induire une baisse du rappel.

4 Conclusion et perspectives

Nous avons présenté une méthode non supervisée d'acquisition de familles morphologiques. Malgré la simplicité de la méthode, les résultats obtenus sont très bons, notamment en terme de précision. L'analyse des résultats montre que les liens morphologiques découverts sont variés : flexion, dérivation et composition. De plus, l'apprentissage est effectué uniquement à partir d'une liste de mots et n'utilise aucune ressource externe. L'approche peut donc être directement appliquée à des langues et des domaines différents, à condition que les mots soient formés par concaténation linéaire de morphèmes.

Des évaluations complémentaires sont toutefois nécessaires. Nous n'avons pour l'heure testé le système que pour du vocabulaire issu de corpus de spécialité en français et en anglais. Il serait intéressant d'évaluer les performances pour d'autres langues plus complexes et pour du vocabulaire non technique. L'utilisation des préfixes aux deux premières étapes de la classification suppose que la langue traitée utilise ce procédé de formation. Or ce procédé n'est pas présent dans toutes les langues : le turc par exemple n'emploie que très peu de préfixes, ce qui rend les deux premières étapes du traitement inutiles. Reste alors à déterminer si le système est capable de produire des regroupements pertinents en utilisant uniquement les suffixes. On peut se poser une question similaire pour le vocabulaire moins technique, où le procédé de préfixation est utilisé moins fréquemment. Des expérimentations complémentaires pourront nous permettre de répondre à ces questions.

Les perspectives d'améliorations du système sont diverses. En effet, le système procède à une classification hiérarchique ascendante stricte, sans parenté multiple et donc sans possibilité pour un mot d'appartenir à deux voire à plusieurs familles différentes. Ceci est souhaitable pour les mots composés, qui font partie de plusieurs familles morphologiques. Il faudrait donc recourir à une forme de classification « floue ». Le système ne permet pas non plus la découverte de nouveaux préfixes, en complément de ceux injectés dans le système lors de la phase d'initialisation. On pourrait envisager d'appliquer une phase de bootstrapping similaire à celle qui permet la découverte de nouvelles signatures. De plus, il serait pertinent d'utiliser la fréquence des signatures au cours de la classification, afin de procéder aux regroupements correspondant aux signatures les plus fréquentes. Les informations contextuelles, disponibles dans les corpus, pourraient également permettre d'améliorer encore les résultats, notamment en terme de précision en validant le fusionnement de deux familles en fonction de la similarité de leurs contextes d'occurrence.

Les perspectives applicatives directement envisageables concernent la recherche d'information et la classification de documents. En recherche d'information, les familles morphologiques peuvent être utilisées pour l'extension de requêtes (Moreau & Claveau, 2006). Pour la catégorisation de documents, les familles peuvent servir de descripteurs des documents à classer (Witschel & Biemann, 2006).

Références

Baayen R. H., Piepenbrock R. & Gulikers L. (1995). *The Celex Lexical Database (Release 2) [CD-ROM].* Philadelphia, PA : Linguistic Data Consortium.

Baroni M. & Bernardini S. (2004). BootCaT : Bootstrapping Corpora and Terms from the Web. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa & R. Silva,

Eds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, p. 1313–1316, Lisbon, Portugal.

Bernhard D. (2006). Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment. In M. Kurimo, M. Creutz & K. Lagus, Eds., *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, p. 19–23, Venice, Italy.

Buitelaar P. & Sacaleanu B. (2002). Extending Synsets with Medical Terms. In *Proceedings of the First International WordNet Conference*, Mysore, India.

Creutz M. & Lagus K. (2002). Unsupervised Discovery of Morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, p. 21–30.

Creutz M. & Lagus K. (2005). Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland.

Gaussier E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Workshop on Unsupervised Methods in Natural Language Processing* : University of Maryland.

Goldsmith J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, **27**(2), 153–198.

Grabar N. & Zweigenbaum P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In P. Amsili, Ed., *Actes de TALN 1999*, p. 175–184, Cargèse.

Hathout N. (2005). Exploiter la structure analogique du lexique construit : une approche computationnelle. *Cahiers de Lexicologie*, **87**(2).

Lepage Y. (1998). Solving analogies on words : an algorithm. In *Proceedings of the 17th international conference on Computational Linguistics*, volume 1, p. 728–734, Morristown, NJ, USA : Association for Computational Linguistics.

Moreau F. & Claveau V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. In *Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, p. 181–192.

Schone P. & Jurafsky D. (2000). Knowledge-Free Induction of Morphology Using Latent Semantic Analysis. In *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, Lisbon, Portugal.

Schone P. & Jurafsky D. (2001). Knowledge-Free Induction of Inflectional Morphologies. In *Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics*, p. 1–9.

Schwab D., Lafourcade M. & Prince V. (2005). Extraction semi-supervisée de couples d'antonymes grâce à leur morphologie. In *Actes de TALN 2005*, p. 73–82.

Witschel H. F. & Biemann C. (2006). Rigorous dimensionality reduction through linguistically motivated feature selection for text categorization. In S. Werner, Ed., *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, volume 1, p. 197–204, Joensuu, Finland.

Zweigenbaum P. & Grabar N. (2000). Liens morphologiques et structuration de terminologie. In *Actes de IC 2000 : Ingénierie des Connaissances*, p. 325–334.

Session

Discours

Enchaînements verbaux – étude sur le temps et l'aspect utilisant des techniques d'apprentissage non supervisé

Catherine RECANATI, Nicoleta ROGOVSCHI
LIPN – UMR 7030 du CNRS, Institut Galilée, Université Paris 13
99, avenue J-B. Clément, 93430 Villetaneuse, France
{Catherine.Recanati,Nicoleta.Rogovschi}@lipn.univ-paris13.fr

Résumé. L'apprentissage non supervisé permet la découverte de catégories initialement inconnues. Les techniques actuelles permettent d'explorer des séquences de phénomènes alors qu'on a tendance à se focaliser sur l'analyse de phénomènes isolés ou sur la relation entre deux phénomènes. Elles offrent ainsi de précieux outils pour l'analyse de données organisées en séquences, et en particulier, pour la découverte de structures textuelles. Nous présentons ici les résultats d'une première tentative de les utiliser pour inspecter les suites de verbes provenant de phrases de récits d'accident de la route. Les verbes étaient encodés comme paires (*cat*, *temps*), où *cat* représente la catégorie aspectuelle d'un verbe, et *temps* son temps grammatical. L'analyse, basée sur une approche originale, a fourni une classification des enchaînements de deux verbes successifs en quatre groupes permettant de segmenter les textes. Nous donnons ici une interprétation de ces groupes à partir de statistiques sur des annotations sémantiques indépendantes.

Abstract. Unsupervised learning allows the discovery of initially unknown categories. Current techniques make it possible to explore sequences of phenomena whereas one tends to focus on the analysis of isolated phenomena or on the relation between two phenomena. They offer thus invaluable tools for the analysis of sequential data, and in particular, for the discovery of textual structures. We report here the results of a first attempt at using them for inspecting sequences of verbs coming from sentences of French accounts of road accidents. Verbs were encoded as pairs (*cat*, *tense*) – where *cat* is the aspectual category of a verb, and *tense* its grammatical tense. The analysis, based on an original approach, provided a classification of the links between two successive verbs into four distinct groups (clusters) allowing texts segmentation. We give here an interpretation of these clusters by using statistics on semantic annotations independent of the training process.

Mots-clés : temps, aspect, sémantique, apprentissage non supervisé, fouille de données.

Keywords: time, tense, aspect, semantics, unsupervised learning, data mining.

1 Introduction

L'intérêt de l'apprentissage *non supervisé* est qu'il permet la découverte de catégories initialement inconnues. Les techniques actuelles permettent d'explorer des séquences de

phénomènes alors qu'on a tendance à se focaliser sur l'analyse de phénomènes isolés ou sur la relation entre deux phénomènes. Elles offrent ainsi de précieux outils pour l'analyse de données organisées en séquences, et en particulier, pour la découverte de structures textuelles. Notre objectif est de les utiliser à cette fin et nous présentons ici l'interprétation des résultats d'une première tentative.

De nombreuses études ont montré l'importance des temps dans la structure narrative d'un récit (Vuillaume, 1990). L'opposition entre le passé simple et l'imparfait a en particulier fait couler beaucoup d'encre. Mais de nombreux liens unissant le temps et l'aspect, l'idée d'un couplage temps et catégorie aspectuelle nous a paru naturelle et intéressante. Il a en effet été démontré qu'on ne peut effectuer l'analyse temporelle des suites d'événements à partir du seul temps grammatical sans faire intervenir l'aspect (voir Kamp H., Vet C. ou Vlach F. dans (Martin et Nef, 1981), (Vet, 1994), (Gosselin, 1996), etc.). On trouve aussi dans la littérature des liens entre l'aspect et d'autres phénomènes sémantiques, comme l'intentionnalité ou la causalité. Dans cette première étude, nous avons cherché à voir si l'on pouvait détecter une certaine régularité dans les enchaînements de verbes au sein des phrases en couplant temps et catégorie aspectuelle des verbes, et s'il était possible, dans un cadre restreint, de leur attribuer un « sens ».

Les récits qui ont été analysés sont des récits d'accident de la route provenant de la partie « observations » d'un constat à l'amiable destiné aux assureurs. Ils nous ont gracieusement été fournis par la MAIF, que nous remercions. Assez courts, leur intérêt *a priori* est de montrer comment s'exprime un accident, ses causes et la responsabilité de ses auteurs, dans un espace limité. Ces textes ont déjà été utilisés au LIPN pour des travaux sur les inférences causales (Nouioua et Kayser, 2006). Notre approche est néanmoins différente puisqu'il s'agit d'une approche statistique visant à catégoriser des enchaînements de verbes – le pari étant que, si de telles classes d'enchaînements existent, elles aient globalement un sens, à tout le moins pour le type de récit considéré.

Nous avons pleinement conscience de la difficulté concernant l'évaluation de ce premier travail, ce dernier étant basé sur le postulat que de tels enchaînements (ici relativement pauvres du point de vue syntaxico-sémantique) puissent se voir attribuer un sens, et ne pas être le reflet de statistiques contingentes. Mais le peu de ressources utilisées est un précieux avantage pour de futures applications au TAL, et l'expérience méritait donc d'être menée. Ajoutons que les outils mathématiques dont nous disposons nous ont permis de tester la validité des catégories obtenues du point de vue statistique, et que notre analyse sémantique a été effectuée à partir d'annotations complètement indépendantes de l'apprentissage.

1.1 Intérêts de notre approche formelle

Les SOM (Self Organizing Map) ou cartes topologiques de Kohonen (Kohonen, 1995) permettent un apprentissage non supervisé (*clustering*) efficace avec visualisation simultanée des résultats de la classification. Cette visualisation se fait grâce à la carte topologique des données (deux données similaires sont proches sur la carte) qui fournit en même temps un codage "intelligent" des données sous forme de prototypes. Ces prototypes étant de même nature que les données, ils sont interprétables, et la carte fournit ainsi un résumé des données. A partir de ce codage, nous avons pris les HMM (Hidden Markov Models) pour modéliser la dynamique des séquences de données (ici, les suites de verbes). Les HMM **(**Rabiner et Juang, 1986) sont la meilleure approche pour traiter des séquences de longueur variable et capturer leur *dynamique*. C'est pourquoi ces modèles ont été largement utilisés dans le domaine de la

reconnaissance de la parole et sont tout particulièrement adaptés à notre objectif. Pour la validation de notre approche, nous avons utilisé à la fois des données génétiques et des données textuelles (celles dont nous présentons ici l'interprétation). Pour plus de détails techniques sur cette méthode, voir (Rogovschi, 2006) ou (Rogovschi et al., 2006).

1.2 Codage des phrases

L'analyse a été réalisée sur une centaine de textes correspondant à 700 occurrences de verbes. On a considéré dans ces récits toutes les séquences de verbes délimitées par des phrases d'au moins deux verbes. Pour palier au faible nombre de données, nous avons utilisé des techniques de ré-échantillonnage basées sur des fenêtres glissantes permettant d'augmenter la redondance (la redondance assure une meilleure classification des données). Pour le codage, les quatre catégories aspectuelles de verbes (état, activité, accomplissement, achèvement) originellement dues à Vendler et Kenny (Vendler, 1967) ont été couplées avec le temps grammatical. Le Tableau 1 résume sommairement les différences entre ces quatre catégories sémantiques. L'indexation a été réalisée à la main en s'appuyant sur notre conception de ces catégories (Recanati C., Recanati F., 1999).

ETAT homogène, duratif, habituel ou indiquant une disposition *être (à l'arrêt) / vouloir / pouvoir*	**ACTIVITE** processus relativement homogène, non borné *rouler / circuler / slalomer / suivre*
ACCOMPLISSEMENT processus dirigé et borné par une fin *traverser / faire un créneau / aller à*	**ACHEVEMENT** événement quasi ponctuel *franchir / heurter / percuter*

Tableau 1 : Les quatre catégories aspectuelles de verbes

Ce type de récit n'utilise globalement que l'imparfait (24%) et le passé composé (34%), avec de temps à autre quelques phrases au présent. On y trouve aussi quelques occurrences (rares) de passé simple et de plus-que-parfait. Il existe par contre un nombre important de participes présents (11%) et d'infinitifs (20%). Nous avons donc décidé de les retenir, bien qu'ils ne participent pas de la même manière à l'ossature grammaticale. Nous avons ainsi effectué l'apprentissage en gardant 9 codes[1] pour les temps apparus sur les verbes.

Exemple « Le véhicule B *circulait* sur la voie de gauche des véhicules *allant* à gauche (marquage au sol par des flèches). Celui-ci *s'est rabattu* sur mon véhicule, me *heurtant* à l'arrière. Il *a accroché* mon pare-choc et m*'a entraîné* vers le mur amovible du pont de Gennevilliers que j*'ai percuté* violemment. » sera réduit après codage aux suites de verbes apparus dans les phrases : (circulait, allant) / (s'est rabattu, heurtant)/ (a accroché, a entraîné, ai percuté) – lesquelles ont encore été encodées numériquement comme séquences de couples (temps, catégorie), soit ici : (act., IM) (acco., pp) / (acco., PC) (achè., pp) / (achè., PC) (acco., PC) (achè., PC).

[1] IM = imparfait, PR = présent, PC = passé composé, PS = passé simple, PQP = plus-que-parfait, inf = infinitif, ppr = participe présent, pp = participe passé et pps = participe passé surcomposé. En ce qui concerne les participes passés (peu nombreux ici), nous n'avons pas toujours compté les emplois adjectivaux contingents, et du point de vue du catégorisateur, ils constituent plutôt du bruit.

2 Premiers résultats

Les premiers résultats concernent les statistiques sur les verbes et les catégories, indépendamment de leurs enchaînements. Dans ce type de récit, les verbes d'état représentent 24% du corpus, ceux d'activité seulement 10%, et l'on trouve 34% de verbes d'accomplissement et 32% de verbes d'achèvement. La répartition non uniforme des temps sur les catégories confirme l'intérêt de notre couplage temps/catégorie aspectuelle (cf. Figure 1).

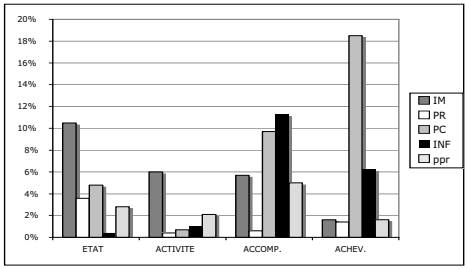


Figure 1 : Répartition des temps par catégorie

Ces pourcentages et leur répartition s'expliquent assez naturellement par la nature de chaque catégorie aspectuelle, la structure généralement typique de ces récits, et la spécialisation aspectuelle des temps grammaticaux (opposition perfectif/imperfectif).

Les verbes d'états (24%) sont répartis à plus de 70% sur l'imparfait, le présent et le participe présent. Cela n'est guère surprenant puisque les états sont homogènes, souvent duratifs ou caractérisant une aptitude (habituels, génériques). La proportion néanmoins non négligeable du passé composé s'explique sans doute par la fréquence de verbes comme « vouloir » ou « pouvoir », classés comme verbes d'états du fait de leur aspect dispositionnel (« j'ai voulu freiner », « je n'ai pu éviter »). La faible proportion de présent provient du fait que le récit est au passé et que le présent historique est trop littéraire pour le genre.

Les verbes d'activités (10%) dénotant des processus homogènes et non bornés, ils se répartissent tout naturellement à plus de 79% sur l'imparfait et le participe présent. La présence de 10% d'infinitif peut facilement s'expliquer par le fait qu'il s'agit de processus qui ont un début, et qui peuvent donc se trouver complément de verbe comme « commencer à », « vouloir », ou être introduits pour mentionner un but par la préposition « pour ».

Les accomplissements (34%) et les achèvements (32%) sont très fréquents au passé composé du fait de leur caractère télique (borné par une fin). Les achèvements sont présents de manière massive car, étant ponctuels ou de courte durée, ils supportent mal l'imparfait. A l'inverse, les accomplissements supportent bien l'imparfait et le participe présent, parce qu'ils ont une durée intrinsèque, et mettent l'accent sur le procès plutôt que sur sa fin – ce qui les rapproche finalement des activités. L'importance globale de ces deux catégories est sans doute liée à la typologie des textes analysés, un récit d'accident impliquant de décrire la séquence des événements successifs qui l'ont provoqué.

Distinction perfectif/imperfectif. Il y a trois points de vue dans le système aspectuel du Français (Smith, 1991). Un point de vue perfectif présente les situations comme fermées, y

compris les états (le point final est alors un changement d'état). Le perfectif s'exprime par le passé composé et le passé simple. Les points de vue imperfectif et neutre présentent à l'inverse des situations ouvertes. Le point de vue neutre s'exprime par le présent. Le point de vue imperfectif s'exprime par l'imparfait, ou par la forme *en train de.* Paul J. Hopper (Hopper, 1979) a fort bien décrit les caractéristiques de ces deux modes vis-à-vis de la structure narrative, du focus et de l'aspect (cf. Tableau 2). Ici, l'opposition perfectif/imperfectif sera réalisée par l'opposition passé composé/imparfait. Cette opposition est néanmoins globale et l'on aurait tort d'attribuer de manière systématique un aspect imperfectif à tous les imparfaits.

Perfectif	**Imperfectif**
Chronologie stricte	Simultanéité ou recouvrement
Vue d'un événement comme un tout, dont la fin est une condition préalable pour l'événement suivant	Vue de la situation ou de ce qui arrive sans que la fin soit nécessaire pour ce qui va arriver ensuite
Identité du sujet à l'intérieur de chaque épisode discret	Changement fréquent de sujet
Topicalisation humaine	Variété de topiques, y compris des phénomènes naturels
Focus non marqué dans les clauses, avec présupposition du sujet et assertion dans le verbe et ses compléments immédiats	Distribution marquée du focus, sur le sujet, l'instrument, ou un adverbe
Evénements dynamiques et cinématiques	Situation descriptive, statique
Avant-plan. Evénement indispensable à la structure narrative	Arrière-plan. Etat ou situation nécessaire pour comprendre les motifs, les attitudes, etc.

Tableau 2 : Opposition perfectif/imperfectif

Structure typique. Un récit d'accident commence généralement par quelques phrases décrivant les circonstances et les états de choses précédant l'accident. Cette première partie est alors à l'imparfait, et contient de nombreux participes présents. On y trouve aussi quelques présents et de nombreux infinitifs introduits par « pour », ou compléments de verbes (« je m'apprêtais à tourner », « le feu venait de passer au rouge »). Essentiellement circonstancielle, cette partie contient une majorité de verbes d'états, et quelques activités et accomplissements. Elle est globalement caractérisée par un point de vue imperfectif, et le récit est en arrière-plan. Vient ensuite la description de l'accident proprement dit, qui mentionne la suite des événements ayant conduit à l'accident pour finir par le choc. Cette partie utilise massivement des verbes d'accomplissement et d'achèvement, le plus souvent au passé composé. Elle est caractérisée par un mode perfectif, mais le but du jeu étant d'indiquer les responsabilités des auteurs, on trouve ici encore beaucoup de participes présents et de tournures infinitives enchaînant souvent trois verbes (« J'ai voulu freiner pour l'éviter », « voulant éviter la borne, je n'ai pu »). En fin de récit, on trouve parfois une troisième partie constituée de commentaires et inventoriant notamment les dégâts. Cette partie est relativement courte et plus difficile à caractériser stylistiquement.

3 Catégorisation des enchaînements verbaux

Notre approche non supervisée a fourni une classification des paires de deux verbes successifs (au sein d'une même phrase) en quatre groupes. Il faut souligner que ce nombre de

quatre, particulièrement petit, est intéressant car il atteste du bien fondé de notre couplage temps/aspect (pas d'explosion combinatoire). Rappelons qu'avec 9 temps et 4 catégories on obtient 36 sortes de verbes, soit 1296 couples virtuels. La répartition des temps sur les catégories a en effet restreint le nombre de couples. Mais c'est aussi la capacité de réduction de la méthode qui permet d'obtenir ce résultat. Une première réduction de dimension a été effectuée par les cartes SOM, qui représentent ici les classes avec 36 paires, puis un élagage de la carte, effectué à partir des matrices de probabilités de transitions provenant des HMM, a réduit encore ces classes à un plus faible nombre de paires typiques.

Annotations sémantiques. Pour faciliter l'interprétation des classes obtenues, nous avions effectué par avance un certain nombre d'annotations. Ces annotations n'ont pas été utilisées pour l'apprentissage mais elles vont nous permettre de caractériser sémantiquement les classes de transitions verbales de façon globale. Ainsi, pour rendre compte de la structure typique de ces récits, nous avons indexé tous les verbes d'un numéro indiquant la « partie » thématique (1-*circonstance*, 2-*accident* ou 3-*commentaire*). Nous avons également marqué certains verbes des attributs *foreground* ou *background* pour indiquer que le récit est en avant-plan ou en arrière-plan. Pour déceler d'éventuelles chaînes causales conduisant à l'accident, nous avons marqué les verbes des attributs *causal* ou *choc* quand le verbe indiquait une cause directe de l'accident, ou le choc lui-même. Nous avons également marqué les verbes d'action en fonction de l'agent (*A* pour le conducteur auteur du récit, *B* pour « l'adversaire », et *C* pour un tiers). Nous avons aussi noté la présence de négation, et l'évocation plus générale de buts poursuivis ou de mondes possibles voisins qui ne se sont pas produits (attributs *négation* et *inertia*). Le Tableau 3 résume symboliquement les résultats que nous avons obtenus en faisant des statistiques sur nos marques sémantiques dans ces quatre classes. Le marquage de la négation s'est avéré peu discriminant, et celui des agents relativement peu informatif.

Groupe C (circonstances)	**Groupe CI (circonstances ou incident)**
Pas de causalité, arrière-plan, pas de choc, nombreux buts et alternatives *(étais, tournant) (reculais, repartir)*	Peu de causalité, arrière-plan, peu de choc, ni but ou alternative *(démarrais, ai entendu) (avait, trouvais)*
Groupe AA (actions menant à l'accident)	**Groupe CC (choc ou commentaires)**
Causalité forte, relief neutre, quelques chocs, nombreux buts et alternatives *(ai voulu, engager) (a percuté, abîmant)*	Causalité très forte, en avant-plan, choc fréquent, buts et alternatives *(a accroché, a entraîné) (n'a pu, stopper)*

Tableau 3 : Synthèse sur les annotations sémantiques

3.1 Groupe C des circonstances

Ce groupe se distingue par une nette différenciation du premier verbe et du second. Le premier verbe est à 93% à l'imparfait, pour seulement 7% de présent, tandis que le second est à 63% à l'infinitif et à 30% au participe présent. Du point de vue des catégories aspectuelles, le premier verbe est à 56% un verbe d'état, et le second à 63% un verbe d'accomplissement (les autres catégories se trouvant distribuées de manière régulière entre 12% et 16%). On pourrait résumer globalement les transitions de ce groupe comme présentant un verbe d'état (ou d'activité) à l'imparfait, suivi d'un verbe d'accomplissement à l'infinitif ou au participe

présent. Le Tableau 4 (voir plus loin) nous donne une synthèse plus fine. Ce groupe privilégie les états et les activités au détriment des achèvements – et les accomplissements sont massivement représentés en seconde occurrence verbale. Ce groupe est celui où l'attribut *inertia* (indiquant un but ou un monde possible proche) est le plus important. Cela s'explique par les nombreux accomplissements introduits par la préposition « pour » (« je *reculais* pour *repartir* », « je *sortais* du parking pour me *diriger* ») ou les auxiliaires à l'imparfait introduisant un infinitif et indiquant des intentions du conducteur (« je m*'apprêtais* à *tourner* à gauche »). C'est une des raisons pour laquelle nous l'avons baptisé groupe C des circonstances. L'autre raison est que ce groupe contient une majorité de verbes appartenant à la première partie du récit (63%), et peu de la seconde et la troisième. On constate en outre que ce groupe ne contient pratiquement aucun verbe indiquant les causes de l'accident ou le choc. L'acteur A y est le plus présent, et le récit est en arrière-plan.

3.2 Groupe CI des circonstances ou de la survenue d'un incident

Le groupe CI est celui des circonstances ou de la survenue d'un incident. On note ici un grand nombre de verbes d'états (37,5%) et d'activités (17%), encore plus important que dans le groupe précédent et très supérieur à la moyenne. On a par contre un nombre moyen d'achèvements (29%), absents du premier verbe mais massivement représentés en second. Cela distingue ce groupe du précédent, où les accomplissements jouaient ce rôle. Ici, à l'inverse, les accomplissements sont exclus de la seconde place et nettement sous représentés (16,5%). On peut synthétiser les enchaînements de ce groupe en disant qu'on a généralement affaire à un état ou une activité à l'imparfait, suivi d'un état ou d'un achèvement, à l'imparfait ou au passé composé. On enchaîne donc deux points de vue imperfectifs, et parfois, un point de vue imperfectif et un point de vue perfectif. On a en effet 36% des verbes qui proviennent de la partie circonstancielle (« Je *circulais* à environ 45 Km/h dans une petite rue à sens unique où *stationnaient* des voitures de chaque côté »), mais également des séquences finissant par un achèvement au passé composé, provenant de la seconde (34%, « Je *roulais* dans la rue Pasteur quand une voiture *a surgi* de ma droite »). Ce groupe contient en outre 25% de séquences verbales situées à cheval entre les deux parties, soit environ la moitié de ces dernières. C'est la raison pour laquelle nous l'avons baptisé « groupe des circonstances ou de la survenue d'un incident ». Le récit est principalement en arrière-plan. L'acteur A (ou un tiers C) se trouvent fortement représentés au détriment de l'acteur B. Il y a peu d'allusions aux causes de l'accident et au choc. L'évocation de buts ou d'alternatives y est insignifiante.

3.3 Groupe AA des actions menant à l'accident

Les verbes du groupe AA proviennent essentiellement du récit de l'accident proprement dit. Ce groupe est caractérisé par l'abondance des accomplissements, au détriment des états et des activités, et c'est pourquoi nous l'avons baptisé « groupe des actions menant à l'accident ». Le mode est généralement perfectif mais on y trouve aussi beaucoup d'infinitifs. Les éléments les plus typiques sont listés sur le Tableau 4. Ces enchaînements se prêtent à des constructions de trois verbes comme « j'*ai voulu m'engager* pour *laisser* », ou « n'*ayant* pas la possibilité de *changer* de voie et la route *étant* mouillée ». 56% des séquences proviennent de la seconde partie, mais les participes présents et les infinitifs permettant l'expression de buts et de mondes possibles (« *désirant* me *rendre* à », « *commençant* à *tourner* »), 26% proviennent de la première partie. On constate ici une assez forte proportion d'acteurs A et B, peu de tiers C, et très peu de marques de relief – le récit n'étant ni spécialement en avant-plan, ni

spécialement en arrière-plan. On trouve beaucoup de verbes participant à la chaîne causale de l'accident, mais relativement peu mentionnant le choc.

3.4 Groupe CC du choc ou des commentaires

Les verbes d'achèvements figurent ici (45%) en plus grand nombre que partout ailleurs, au détriment des activités (seulement 6,5%) et des états (seulement 14,5%). Cela explique que ce groupe favorise globalement la partie descriptive de l'accident (57%). On observe aussi une augmentation des infinitifs et des participes en premier verbe au détriment de l'imparfait et du présent, et une augmentation massive du passé composé sur le second verbe au détriment de toutes les catégories – sauf le présent (8%, légèrement plus que la moyenne). Cette apparition du présent explique peut-être la présence de la partie 3-commentaire (29% au lieu de18% en moyenne). La mention de but ou d'alternative est moyenne. C'est ici par contre que l'avant-plan est le plus marqué. Il y a un nombre important d'acteur B (le conducteur adverse) et c'est là que l'on trouve le plus de verbes relatifs aux causes de l'accident et au choc lui-même. Le Tableau 4 indique seulement deux éléments typiques dans ce groupe qui, bien qu'assez volumineux, est plus difficile à caractériser. L'analyse en termes de point de vue montre néanmoins que la séquence finit généralement par un point de vue perfectif.

Groupe	Type	verbe 1	verbe 2
C	1	Etat ou act., IM	Etat ou act., ppr
	2	Etat, IM (ou PR)	Acc., INF
	3	Act. ou ach., IM	Acc. (ou ach.) INF
CI	4	Etat ou act., IM	Etat (ou ach.), IM
	5	Etat ou act., IM	Etat (ou ach.), PC
AA	6	Acc.(ou ach.) INF	Acc.(ou ach.) INF (ou ppr)
	7	Ach. (ou acc.), PC	Acc.(ou ach.) INF
	8	Etats, PC	Ach. INF
CC	9	Ach. (ou acc.)., INF	Ach., PC
	10	Ach.ou état, PC	Ach. (ou acc) PC

Tableau 4 : Éléments les plus typiques des quatre groupes

3.5 Bilan et commentaires

Cette catégorisation a bien distingué les états et activités (groupes C, CI) des événements (groupes AA, CC). De manière plus intéressante, les accomplissements sont aussi distingués des achèvements, justifiant la distinction accomplissement/achèvement (par opposition à la notion plus générale d'événements). On a pu également mettre en évidence que l'expression de buts ou d'alternatives passe souvent par l'utilisation de verbes au participe présent ou à l'infinitif – ce qui explique les taux réalisés par les groupes C et AA. Mais la catégorie utilisée influence aussi cette expression, car le second verbe dans ces deux groupes est généralement un accomplissement. En outre, les groupes C et CI (non marqué pour cet indice) se distinguent justement sur le type d'événement qui apparaît en second. De même les éléments différenciant les groupes AA et CC (lesquels ont cette fois une majorité d'événements) montrent que le groupe AA (qui favorise les accomplissements), bien que véhiculant un mode perfectif, est peu marqué sur le plan du relief narratif. Ce groupe est également moins concerné par les causes de l'accident que le groupe CC, et il fait peu allusion au choc. Les

buts et intentions s'exprimeraient donc plus facilement par des accomplissements que par des achèvements – lesquels seraient porteurs de plus de causalité. En effet, le groupe CC, qui favorise les achèvements, est plus fortement marqué pour l'avant plan, le choc lui-même et la chaîne causale des événements l'ayant directement provoqué. Ajoutons à ce propos qu'il semble que pour les achèvements et les activités, le sujet ait une relation de pouvoir sur l'objet direct (ou sur l'objet oblique). On peut tester son existence en utilisant des adverbes de manière (doucement, précautionneusement, etc.). Cela explique peut-être aussi la plus forte responsabilité du sujet avec des verbes d'achèvements.

Mais quoi qu'ayant bien repéré l'opposition perfectif/imperfectif (groupes AA et CC vs C et CI), cette classification a mis dans le même groupe CI des séquences à l'imparfait et les ruptures imparfait/passé composé. Une des explications est que notre algorithme d'apprentissage n'a pas tenu compte de l'ordre des phrases (ni de la distinction entre les textes), de sorte que la succession de plusieurs phrases à l'imparfait, et la structure typique de ces récits (telle que nous la percevons), n'a pas pu être bien repérée. On a ainsi manqué une part importante de notre objectif. Mais les résultats obtenus sont déjà prometteurs, puisque les trois parties ont tout de même été distribuées de manière non uniforme sur les quatre groupes. On notera également que cette classification a mis en évidence l'importance des tournures infinitives et des participes présents, et la subtilité de leurs enchaînements (cf. Tableau 4).

Améliorations techniques possibles. On a construit ici les HMM en déplaçant une fenêtre de taille 2 : un verbe est analysé au regard du verbe qui le précède et de celui qui le suit, mais pas au regard des *n* précédents ou des *n* suivants. Cela n'est pas très gênant si l'on se situe comme ici au niveau de la phrase, (dans ces récits, les phrases comportent rarement plus de trois verbes) mais pour une analyse globale prenant en compte tout le texte, on aura certainement besoin de cette amélioration. D'autre part, nous aurions aimé produire des séquences typiques de longueur variables. Ainsi, les groupes AA et CC auraient fourni des séquences de plusieurs verbes. (Cela se voit sur les éléments du Tableau 4, et nous l'avons par ailleurs constaté sur la segmentation des textes). Ce résultat pourrait être obtenu automatiquement à partir des HMM, mais nous n'avons pas eu encore le temps d'implanter cette méthode.

4 Conclusion

Notre projet général est d'appliquer les techniques de la fouille de données à la découverte de structures textuelles. Nous avons développé à cette fin une technique d'apprentissage non supervisé permettant de détecter des structures séquentielles. Elle a permis d'analyser les séquences de verbes constituant une phrase, et de proposer une classification des apparitions de deux verbes successifs en quatre groupes. Nous avons réussi à valider sémantiquement ces groupes de manière satisfaisante, en nous basant sur des annotations et des statistiques. Cela confirme à la fois le bien-fondé de la technique employée, et celui de notre couplage des temps grammaticaux avec la catégorie aspectuelle d'un verbe.

Mais ce travail n'en est encore qu'à ses débuts, et il nous reste de nombreux points à élucider. Nous regrettons tout d'abord de ne pas avoir pu comparer nos statistiques globales sur les emplois temps/catégorie à celles d'autres types de récits (et en particulier à celle de récits simples d'incident au passé). Il nous reste en effet à déterminer quelle est la part "typologique" des groupes d'enchaînements que nous avons isolés. Nous n'avons pas non plus eu le temps d'exploiter les automates probabilistes obtenus à partir de notre méthode, et ces derniers pourraient se révéler intéressants pour des applications (en particulier en génération). Il reste enfin des améliorations à apporter à la méthode générale pour prendre en compte la

structure globale des textes (non prise en compte ici), et la modélisation reste à poursuivre pour la recherche de séquences de longueur supérieure à 2.

Remerciements

Nous remercions vivement Y. Bennani pour l'aide précieuse qu'il a apportée pour le développement de la technique utilisée, et sans qui ce travail n'aurait pu être mené à bien. Nous remercions également A. Nazarenko et D. Kayser pour leurs aimables relectures.

Références

GOSSELIN L. (1996), Sémantique de la temporalité en français, Louvain-la-Neuve : Duculot.

HOPPER J. (1979). Some observations on the typology of focus and aspect in narrative language. *Studies in Language* 3.1, 37-64, Amsterdam : J. Benjamins.

KOHONEN T., (1995). *Self-Organizing Map*. Springer.

MARTIN R, NEF F. eds (1981). Le temps grammatical. *Langage* 64, KAMP H. 39-64, VLACH F. 65-79, VET C. 109-124, Paris : Larousse.

NOUIOUA F., KAYSER D. (2006). Une expérience de sémantique inférentielle. *Actes de TALN 2006,* 246-255.

RABINER L.R., JUANG B.H. (1986). An Introduction to Hidden Markov models. *IEEE ASSP Magazine*, jan. 86, 4-16.

RECANATI C., RECANATI F. (1999). La classification de Vendler revue et corrigée. *La modalité sous tous ses aspects, Cahiers Chronos 4*, 167-184. Amsterdam/Atlanta, GA.

ROGOVSCHI N. (2006). *Systèmes d'apprentissage non supervisé connexionnistes et stochastiques pour la fouille de données structurées en séquences*. Rapport de stage de Master Recherche, LIPN, Université Paris 13.

ROGOVSCHI N., BENNANI Y., RECANATI C. (2007). Apprentissage neuro-markovien pour la classification non supervisée de données structurées en séquences. Actes des *7èmes journées francophones Extraction et Gestion des Connaissances*. Namur, Belgique.

SMITH C. S. (1991). *The parameter of aspect, Studies in Linguistics and Philosophy*, Kluwer Academic publishers.

VENDLER Z. (1967). Verbs and Times. *Linguistics in Philosophy*, 97-121. Ithaca, New-York: Cornell University Press.

VUILLAUME M. (1990). *Grammaire temporelle des récits*. Paris : Minuit.

VET C. (1994). Relations temporelles et progression thématique. *Études Cognitives 1, Sémantique des Catégories de l'aspect et du Temps*, 131-149. Warszawa : académie des Sciences de Pologne.

D-STAG : un formalisme pour le discours basé sur les TAG synchrones

Laurence DANLOS
LATTICE – Université Paris 7
Institut Universitaire de France
Laurence.Danlos@linguist.jussieu.fr

Résumé. Nous proposons D-STAG, un formalisme pour le discours qui utilise les TAG synchrones. Les analyses sémantiques produites par D-STAG sont des structures de discours hiérarchiques annotées de relations de discours coordonnantes ou subordonnantes. Elles sont compatibles avec les structures de discours produites tant en RST qu'en SDRT. Les relations de discours coordonnantes et subordonnantes sont modélisées respectivement par les opérations de substitution et d'adjonction introduites en TAG.

Abstract. We propose D-STAG, a framework which uses Synchronous TAG for discourse. D-STAG semantic analyses are hierarchical discourse structures richly annotated with coordinating and subordinating discourse relations. They are compatible both with RST and SDRT discourse structures. Coordinating and subordinating relations are respectively modeled with the TAG substitution and adjunction operations.

Mots-clés : discours, grammaires d'arbres adjoints (synchrones), interface syntaxe/sémantique.

Keywords: discourse, (synchronous) tree adjoining grammars, syntax/semantic interface.

1 Introduction

RST - Rhetorical Structure Theory (Mann & Thompson, 1988; Taboada & Mann, 2006) - et SDRT - Segmented Discourse Representation Theory (Asher, 1993; Asher & Lascarides, 2003) - sont deux théories pour le discours qui reposent sur la notion de *relation de discours*. Ces théories partagent l'idée que certaines parties d'un discours, appelées Satellites, jouent un rôle « subordonné » (« moins important ») que d'autres parties, appelées Nucleus. Cette asymétrie est comparable à la distinction faite au niveau syntaxique entre les arguments (nuclei) et les adjoints (satellites) d'une phrase. Elle amène à poser l'existence de deux types de relations de discours : une relation *coordonnante (multi-nucléaire)* relie deux Nuclei, tandis qu'une relation *subordonnante (nucleus-satellite)* relie un Nucleus et un Satellite. Elle permet la construction de structures de discours hiérarchiques annotées de relations coordonnantes ou subordonnantes.

Nous proposons ici un nouveau formalisme pour le discours, D-STAG (Discourse Synchronous TAG), qui utilise aussi les relations de discours coordonnantes et subordonnantes et qui repose

sur STAG - Synchronous TAG (Shieber, 1994; Shieber & Schabes, 1990). STAG a été utilisé avec succès dans une grammaire anglaise qui permet d'engendrer simultanément les analyses syntaxique et sémantique d'une phrase (Nesson & Shieber, 2006). Dans le prolongement de cette idée, une grammaire D-STAG pour le français ou l'anglais permet d'engendrer simultanément les analyses syntaxique et sémantique d'un discours français ou anglais donné en entrée. Les analyses sémantiques des discours sont des structures hiérarchiques annotées par des relations coordonnantes ou subordonnantes. Elles peuvent être converties de façon déterministe en structures RST ou SDRT. Par conséquent, D-STAG peut bénéficier des résultats apportés par ces théories du discours, par exemple, D-STAG peut bénéficier de la *Contrainte de la Frontière Droite* (Section 2.2), qui simplifie grandement le calcul des structures de discours.

En D-STAG, les relations de discours coordonnantes et subordonnantes sont modélisées respectivement par les opérations de substitution et d'adjonction introduites en TAG- Tree Adjoining Grammar (Joshi, 1985). En TAG, les arbres initiaux introduisent les dépendances prédicat-arguments par substitution, tandis que les arbres auxiliaires introduisent la récursivité et permettent la modification d'arbres élémentaires grâce à l'opération d'adjonction. D-STAG suit ces principes de base de TAG ; plus précisément, les relations coordonnantes ancrent des arbres initiaux qui introduisent leurs arguments (deux Nuclei) par substitution, tandis que les relations subordonnantes ancrent des arbres auxiliaires : le Satellite d'une relation subordonnante modifie son Nucleus par l'adjonction d'un arbre auxiliaire ancré par la relation subordonnante.

Une grammaire D-STAG est une extension naturelle d'une grammaire STAG réalisant une interface syntaxe/sémantique pour les phrases. Dans cette perspective, le même analyseur peut être utilisé tant pour les phrases que pour les discours, et les formes logiques des discours peuvent être calculées de façon déterministe à partir des analyses sémantiques calculées en D-STAG en employant les formes logiques des phrases (qui sont calculées de façon déterministe à partir de leurs analyses sémantiques produites par la grammaire STAG).

D-STAG ressemble à D-LTAG - Discourse Lexicalized TAG dans la version présentée dans (Forbes-Riley *et al.*, 2006) - dans la mesure où les deux formalismes étendent une interface syntaxe/séman-tique phrastique au niveau du discours. Cependant, il existe une différence cruciale : D-LTAG n'utilise pas les relations de discours et ignore la distinction entre relations coordonnantes et subordonnantes. Par conséquent, D-LTAG n'a rien de commun avec RST ou SDRT.

Cet article est organisé de la façon suivante : la Section 2 introduit brièvement RST et SDRT, la Section 3 STAG. La Section 4 expose les principes de D-STAG. La Section 5 compare D-LTAG et D-STAG. Le même exemple de référence est utilisé dans toutes les sections, à savoir (1) qui est de la forme P_1 *parce que* P_2. *Ensuite*, P_3., dans lequel *parce que* exprime la relation subordonnante Explication et *ensuite* la relation cordonnante Narration.

(1) Jean est allé au super-marché parce que son frigo était vide. Ensuite, il est allé au cinéma.

2 Brève introduction à RST et SDRT

RST et SDRT reposent sur la notion de relation de discours et emploient *grosso modo* le même ensemble de relations de discours[1]. Des débats ont lieu sur la distinction Nucleus/Satellite de

[1]Il arrive qu'une relation de discours donnée soit nommée de deux façons différentes. Par exemple, Séquence en RST est appelée Narration en SDRT.

RST (Stede, 2007) et sur la distinction cordonnante/subordonnante de SDRT (Asher & Vieu, 2005). Cependant, on suppose ici que le statut d'une relation de discours (coordonnante versus subordonnante) est clair et identique en RST et SDRT[2]. RST et SDRT proposent des structures de discours graphiquement différentes, comme montré dans les deux sections suivantes.

2.1 Structures de discours en RST

RST est une théorie qui a été conçue il y a une vingtaine d'années et qui a été beaucoup employée. De ce fait, il existe différentes interprétations de RST (Taboada & Mann, 2006). Dans cet article, nous ne parlerons que de l'interprétation de Marcu, qui a eu un fort impact en analyse de discours (Marcu, 2000) et annotation de discours (Carlson *et al.*, 2003).

Pour un discours donné, RST calcule une structure d'arbre qui connecte récursivement les unités minimales (les clauses) et les segments de texte plus larges ainsi construits ; une relation de discours ne peut relier que des segments de texte adjacents. L'arbre RST pour (1), dans la représentation graphique de Marcu (2000), est donné dans la Figure 1(a). Le symbole C_i représente la phrase P_i, son analyse syntaxique ou son analyse sémantique (selon l'application dans laquelle RST est employée). Les étiquettes N et S sur les arcs abrègent respectivement Nucleus et Satellite.

Marcu (2000) a proposé un principe, appelé « Principe de Nucléarité » (ou « Principe compositionnel ») qui s'énonce comme suit : quand une relation de discours relie deux segments de discours, alors elle relie aussi les Nuclei de ces deux segments. Le Principe de Nucléarité donne l'interprétation en termes de prédicats et arguments à un arbre RST. Par exemple, il indique que C_1 est l'argument gauche de Narration dans l'arbre pour (1) donné dans la Figure 1(a).

FIG. 1 – Arbre RST (a) et graphe SDRT (b) pour le discours (1)

2.2 Structures de discours en SDRT

À l'origine, SDRT a été conçue comme une extension de la DRT - Discourse Representation Theory (Kamp & Reyle, 1993) - pour rendre compte des propriétés spécifiques du discours. Toutefois, cette théorie s'est inspirée de RST, par exemple, la distinction coordonnante/subordonnante faite en SDRT est inspirée de la distinction multi-nucléaire/nucleus-satellite faite en RST.

Les structures de discours en SDRT sont représentées comme des graphes dirigés conçus avec les principes suivants. Pour un discours composé de deux phrases (clauses) liées par une relation

[2]Ceci est le cas la plupart du temps, excepté pour Résultat qui est subordonnante en RST et coordonnante en SDRT, bien que (Asher & Vieu, 2005) postulent que Résultat est coordonnante 'par défaut'.

de discours R, les nœuds du graphe sont les étiquettes π_1 et π_2 des DRS donnant les formes logiques des phrases. Ils sont liés par une *flèche* qui est étiquetée par la relation R. La flèche est horizontale si R est coordonnante, verticale si R est subordonnante (Asher & Vieu, 2005). Si l'on prend en compte la distinction Nucleus/Satellite, ceci signifie qu'une flèche horizontale relie deux Nuclei, tandis qu'une flèche verticale descend d'un Nucleus vers un Satellite. En plus des nœuds représentant les phrases (notés π_i et appelés « nœuds de phrase »), les graphes SDRT peuvent comporter des « nœuds de portée » (notés π', π $>>$, ...). Un nœud de portée est lié par des *lignes* à des nœuds de phrase. Pour (1), le graphe SDRT est donné dans la Figure 1(b).

Lors de la procédure de construction des structures de discours, SDRT fait appel à la notion de *frontière droite*, proposée à l'origine par (Polanyi, 1988). Informellement, dans le graphe SDRT pour un discours comportant n phrases (clauses), la frontière droite contient le nœud π_n représentant la dernière phrase et les autres nœuds de phrase situés sur la frontière droite du graphe[3]. A titre d'illustration, la frontière droite dans le graphe de la Figure 1(b) contient seulement le nœud π_3. Lors de la construction dynamique d'un graphe SDRT, les nœuds de la frontière droite sont les seuls nœuds qui permettent d'accrocher une information nouvelle ; ceci est connu sous le nom de « Contrainte de la Frontière Droite »[4]. Cette contrainte simplifie grandement la construction des structures de discours, et par là-même le calcul des formes logiques des discours, qui sont obtenues de façon déterministe à partir des structures de discours.

3 Introduction aux TAG synchrones

Les parties entre guillemets de cette section sont traduites de (Nesson & Shieber, 2006). Les opérations de substitution et adjonction introduites en TAG pour la syntaxe sont rappelées dans le Figure 2.

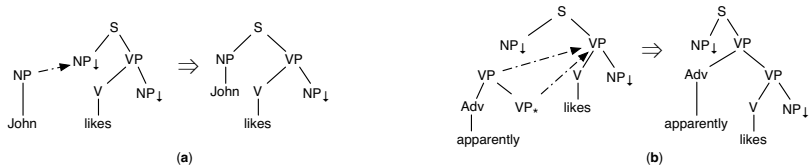


FIG. 2 – Exemples des opérations de substitution (a) et d'adjonction (b) en TAG

« Les TAG synchrones (STAG) prolongent les TAG en considérant les structures élémentaires comme des paires d'arbres TAG avec des liens entre certains nœuds de ces arbres. Une STAG est un ensemble de triplets $\langle t_L, t_R, \frown \rangle$ où t_L et t_R sont des arbres élémentaires TAG et $\frown$ est une relation de liage entre certains nœuds de t_L et certains nœuds de t_R (Shieber, 1994; Shieber & Schabes, 1990). La dérivation se déroule comme en TAG excepté que toutes les opérations doivent être appariées. En d'autres mots, un arbre ne peut être substitué ou adjoint à un nœud que si l'arbre apparié est simultanément substitué ou adjoint au nœud lié. Nous notons les liens en utilisant des indices dans des cercles (e.g. ①) qui viennent décorer les nœuds liés. »

[3]En fait, la frontière droite contient aussi les « nœuds topiques » situés sur la frontière droite du graphe. La notion de nœud topique est importante en SDRT, mais elle est laissée de côté dans cet article.

[4]Cette contrainte dit aussi que l'antécédent d'une expression anaphorique doit être sur la frontière droite, mais les expressions anaphoriques ne sont pas discutées ici.

STAG a été utilisé avec succès dans une interface syntaxe/sémantique pour l'anglais qui peut traiter de phrases complexes soulevant des problèmes délicats de portée (Nesson & Shieber, 2006). Cette interface est illustrée dans la Figure 3 pour l'analyse de la phrase (très simple) *John apparently likes Mary.*

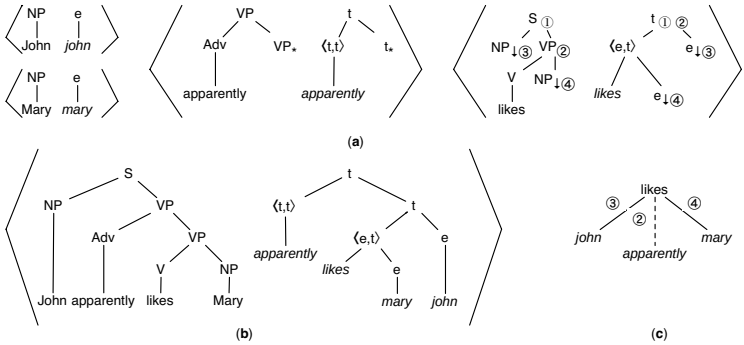


FIG. 3 – Fragment d'une interface syntaxe/sémantique pour l'anglais (a), les paires d'arbres dérivés (b) et l'arbre de dérivation (c) pour la phrase *John apparently likes Mary.* (Figure reproduite à partir de (Nesson et Shieber, 2006))

Dans l'arbre de dérivation, « les opérations de substitution sont notées avec une ligne continue, celles d'adjonction avec une ligne en pointillés. Il faut noter que chaque lien de l'arbre de dérivation spécifie un numéro de liage dans la paire d'arbres élémentaires. Ces liens donnent l'adresse des opérations dans l'arbre syntaxique et dans l'arbre sémantique. Ces opérations doivent opérer à des nœuds liés dans la paire d'arbres élémentaires concernée. »

Les arbres dérivés sémantiques permettent de calculer de façon déterministe les formes logiques des phrases, par exemple la forme $apparently(likes(john, mary))$ à partir de l'arbre dérivé sémantique de la phrase *John apparently likes Mary.*

4 Présentation de D-STAG

D-STAG utilise STAG pour le discours, une grammaire D-STAG étant une extension naturelle d'une grammaire STAG réalisant une interface syntaxe/sémantique au niveau phrastique. Une paire d'arbres élémentaires en D-STAG consiste en un arbre élémentaire ancré par un connecteur de discours et comportant **deux** nœuds non terminaux[5] apparié avec un arbre élémentaire ancré par la relation de discours exprimée par le connecteur et comportant **deux** nœuds non terminaux. Pour une relation de discours coordonnante, l'arbre élémentaire est un arbre initial comportant deux nœuds à substitution correspondant aux deux Nuclei. Pour une relation subordonnante R, il existe deux arbres élémentaires associés aux positions du Satellite vis-à-vis du Nucleus ; la notation R_r (resp. R_l) signifie que le Satellite apparaît à la droite (resp. gauche) du Nucleus.

[5]L'arbre D-STAG d'un connecteur n'est pas forcément similaire à son arbre syntaxique. De ce fait, un analyseur D-STAG doit inclure un module d'extraction d'arbres et un module de correspondance entre arbres syntaxiques et arbres discursifs, comme c'est le cas dans l'analyseur pour D-LTAG décrit dans (Webber, 2004).

Ces arbres élémentaires sont des arbres auxiliaires dont le nœud pied correspond au Nucleus et qui comportent un nœud à substitution correspondant au Satellite.

La Figure 4 contient quatre paires d'arbres : la première, nommée α*ensuite*-Narration[6] apparie l'arbre initial pour *ensuite* avec l'arbre initial pour Narration exprimée par *ensuite*. Les deux suivantes, nommées β*parce_que*-Explication$_R$ et β*parce_que*-Explication$_L$, apparient les arbres auxiliaires pour *parce que* avec ceux pour la relation subordonnante Explication exprimée par *parce que*[7]. La dernière paire d'arbres, nommée α*P-to-D*, est spéciale : elle est conçue pour immerger dans une grammaire D-STAG les analyses syntaxique et sémantique d'une phrase engendrées par une grammaire STAG. Expliquons les symboles non terminaux figurant dans ces paires d'arbres. Les symboles DC et DR sont respectivement utilisés pour les connecteurs de discours (« discourse connectives ») et les relations de discours (« discourse relations ») ; les nœuds étiquetés DC ou DR sont liés par l'indice ③ dont l'emploi sera illustré pour le discours (2) ci-dessous. Les symboles DU et AO sont respectivement utilisés pour les unités de discours (« discourse units ») et les objets abstraits (« abstract objects », (Asher, 1993)). Une unité de discours peut être simple ou complexe. Une unité de discours simple est l'analyse syntaxique d'une phrase « simple » (i.e. une phrase ne comportant pas de connecteur de discours), soit un arbre de racine P qui est introduit au niveau discursif par α*P-to-D*. Une unité de discours complexe est récursivement l'analyse syntaxique d'une phrase complexe (comportant un ou plusieurs connecteurs) ou d'un texte de plusieurs phrases. Parallèlement, les objets abstraits simples ou complexes sont des analyses sémantiques ; un objet abstrait simple est un arbre de racine t qui est introduit au niveau discursif par α*P-to-D*. Signalons que l'étiquetage par N (Nucleus) ou S (Satellite) des arcs pointant sur les nœuds AO dans les arbres élémentaires ancrés par une relation de discours est une information qui est juste destinée à la conversion d'une analyse sémantique de D-STAG en un arbre RST (voir Figure 6(a) ci-dessous).

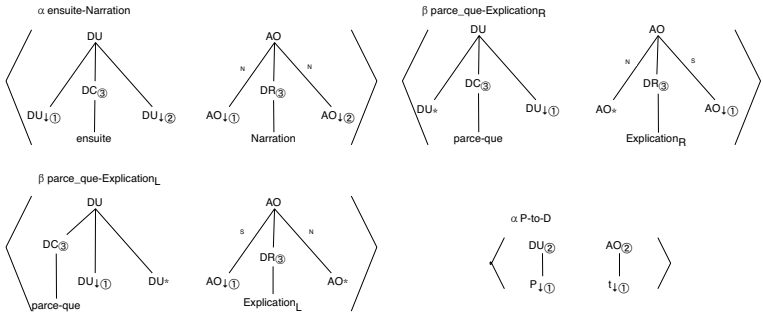


FIG. 4 – Fragment d'une grammaire D-STAG pour le français

Montrons maintenant comment les paires d'arbres de la Figure 4 sont utilisées pour analyser le discours (1). La Figure 5 contient les paires d'arbres dérivés (a)-(b) pour (1), et l'arbre de

[6]Une paire d'arbres élémentaires en D-STAG est nommée avec les conventions suivantes : le préfixe α ou β indique si les arbres appariés sont initiaux (α) ou auxiliaires (β) ; le nom concatène les ancres des arbres élémentaires, i.e un connecteur de discours et une relation de discours.

[7]Dans les phrases de forme P_1 *parce que* P_2. ou *Parce que* P_1, P_2., les nœuds pied proviennent structurellement des phrases principales, les nœuds à substitution des phrases subordonnées. Néanmoins, dans les phrases de forme P_1 *Conj* P_2 *parce que* S_3., dans lesquelles *Conj* désigne une autre conjonction de subordination, le nœud pied de *parce que* n'est pas structurellement défini : il provient de P_1, de P_2 ou de P_1 *Conj* P_2 (Danlos, 2004).

dérivation (c). En supposant que les analyses syntaxique et sémantique d'une phrase simple sont engendrées simultanément par une grammaire STAG, le symbole T_i représente l'analyse syntaxique de la phrase P_i (un arbre de racine P), F_i son analyse sémantique (un arbre de racine t), τ_i son arbre de dérivation. Nous emploierons les termes suivants : dans la Figure 5, (a) est l'*analyse syntaxique discursive* de (1), (b) son *analyse sémantique discursive*, (c) son *analyse sémantique compositionnelle*.

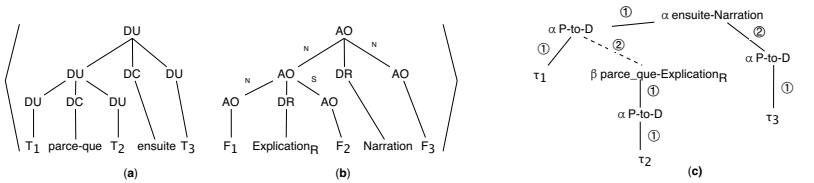


FIG. 5 – Paire d'arbres dérivés (a)-(b) et arbre de dérivation (c) calculés en D-STAG pour (1)

Nous allons maintenant montrer comment les analyses de D-STAG se convertissent en structures de discours RST ou SDRT. Les analyses sémantiques discursives de D-STAG sont convertibles de façon déterministe en arbres RST, par exemple l'analyse (b) de (1) est convertible en l'arbre RST pour (1) - Figure 1(a) - en appliquant récursivement le patron donné dans la Figure 6(a). Les symboles l_1 ou l_2 sur les arcs ont pour valeur N (Nucleus) ou S (Satellite)[8]. La conversion d'analyses sémantiques de D-STAG en arbres RST s'explique en reconsidérant le Principe de Nucléarité (Section 2.1) à la lumière de l'opération d'adjonction. Par exemple, dire que C_1 est l'argument gauche de Narration dans l'arbre RST de la Figure 1(a) (grâce au Principe de Nucléarité) revient à dire que le sous-arbre dont la racine est la relation subordonnante Explication est introduit par **adjonction**. C'est la raison pour laquelle nous avons postulé que les relations de discours subordonnantes ancrent des arbres **auxiliaires** en D-STAG. Comme la distinction Nucleus/Satellite faite en RST est similaire à la distinction argument/adjoint faite en syntaxe, les relations subordonnantes sont modélisées par l'opération d'adjonction, tandis que les relation coordonnantes sont modélisées par l'opération de substitution (Section 1).

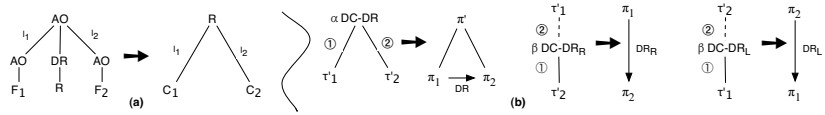


FIG. 6 – Patrons de conversion d'une analyse sémantique de D-STAG vers un arbre RST (a) et d'une analyse compositionnelle de D-STAG vers un graphe SDRT (b)

Passons aux structures de discours SDRT. Les analyses sémantiques compositionnelles de D-STAG sont convertibles de façon déterministe en graphes SDRT, par exemple l'analyse (c) de (1) est convertible selon le graphe SDRT pour (1) - Figure 1(b) - en appliquant récursivement les patrons donnés dans la Figure 6(b). Pour simplifier la lecture de ces patrons, nous avons utilisé les symboles τ_i' qui représentent α*P-to-D* dans lequel τ_i est substitué au lien ①.

[8]Un constituant C_i est soit une phrase, soit son analyse syntaxique, soit son analyse sémantique (alors $C_i = F_i$), selon l'interprétation et l'utilisation de RST (Section 2.1). Dans tous les cas, C_i peut être obtenu à partir de F_i.

Les analyses sémantiques compositionnelles de D-STAG conduisent de façon déterministe aux formes logiques. Par exemple, l'arbre de dérivation (c) pour (1) conduit à la forme logique (simplifiée) suivante : $F_1 \wedge F_2 \wedge F_3 \wedge precede(F_1, F_3) \wedge cause(F_2, F_1)$. En suivant SDRT, ce calcul demande simplement d'interpréter Narration(π_1, π_3) en $F_1 \wedge F_3 \wedge precede(F_1, F_3)$ et Explication(π_1, π_2) en $F_1 \wedge F_2 \wedge cause(F_2, F_1)$.

En conclusion, les analyses sémantiques discursives et compositionnelles de D-STAG sont respectivement convertibles en structures de discours RST et SDRT. Par conséquent, D-STAG peut bénéficier des résultats apportés par ces théories du discours. Par exemple, D-STAG peut profiter de la Contraint de la Frontière Droite (Section 2.2) pour simplifier grandement la construction des structures de discours et par là-même le calcul des formes logiques de discours.

De plus, D-STAG peut profiter de l'opération d'adjonction pour la *modification* des relations de discours, un phénomène qui n'est pas traité en RST ou SDRT. Ce phénomène est illustré par le discours (2).

(2) Tu ne dois pas faire confiance à Jean parce que, par exemple, il ne rend jamais ce qu'il a emprunté. (Exemple traduit de (Webber *et al.*, 2003))

Comme expliqué dans (Webber *et al.*, 2003), l'interprétation de (2) est que le non retour par Jean des objets empruntés est un exemple des raisons pour ne pas lui faire confiance. Guidée par cette interprétation, nous postulons que *par exemple* dans ce discours est un modifieur de *parce que*. De ce fait, nous postulons qu'en D-STAG cet adverbial ancre un arbre auxiliaire syntaxique dont la racine est un nœud étiqueté DC et qui est apparié avec un arbre auxiliaire sémantique dont la racine est étiquetée DR et dont l'ancre est simplement *par-ex'*[9]. La paire d'arbres ainsi formée, appelée βpar-ex, est montrée dans la Figure 7(a). Lors de l'analyse de (2), βpar-ex s'adjoint au lien ③ dans βparce_que-Explanation$_R$ donné dans la Figure 4. L'analyse compositionnelle de (2) est présentée dans la Figure 7(b).

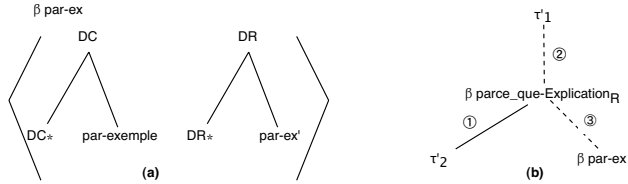


FIG. 7 – Paire βpar-ex (a) et analyse compositionnelle de (2) (b)

La forme logique de (2), e.g. $Exemplify(F_2, \lambda F.cause(F, F_1))$ donnée dans (Forbes-Riley *et al.*, 2006), est calculée simplement à partir de l'arbre de dérivation de (2) en interprétant βpar-ex selon la forme abstraite $Exemplify(?X_2, \lambda F.?R(F, ?X_1))$ dans laquelle les variables $?R$, $?X_1$ et $?X_2$ prennent leur valeur dans l'arbre de dérivation où βpar-ex est adjoint.

Nous concluons cette présentation de D-STAG par une remarque sur l'ambiguïté au niveau discursif. Il arrive souvent qu'un connecteur de discours soit sémantiquement ambigu, i.e. exprime

[9]Par contre, en D-LTAG, *for example (par exemple)* est considéré comme un connecteur de discours en (2). En D-STAG, cet adverbial n'est un connecteur de discours (exprimant la relation de discours Exemplification) que dans un discours comme *Jean adore le fromage. Par exemple, il adore le brie.*

plusieurs relations de discours. C'est le cas entre autres pour le connecteur vide ϵ[10] qui exprime les relations Explication$_r$, Elaboration$_r$ et Narration, entre autres. En D-STAG, un connecteur ambigu ancre autant d'arbres élémentaires syntaxiques qu'il a d'interprétations, ce qui conduit à une paire d'arbres pour chaque interprétation. Le choix de la bonne interprétation pour un connecteur ambigu dépend de considérations (extra)-linguistiques. Par exemple, le discours (3) doit recevoir les analyses présentées dans la Figure 8.

(3) Jean est allé au super-marché parce que son frigo était vide. Il a acheté un rôti et du brie.

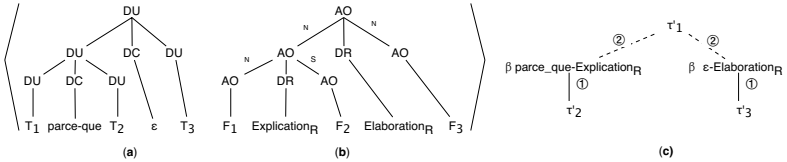


FIG. 8 – Paire d'arbres dérivés (a)-(b) et arbre de dérivation (c) calculés en D-STAG pour (3)

Le fait que le connecteur vide ϵ introduisant P_3 dans (3) exprime la relation Elaboration$_r$ vient de la connaissance qu'on va au super-marché pour faire des achats. La prise en compte de considérations (extra)-linguistiques demande des techniques comme celles mises en œuvre en SDRT, qui reposent sur une procédure incrémentale basée sur la « glue logique » (Asher & Lascarides, 2003). Une autre solution consiste à faire appel à des méthodes probabilistes, comme cela est proposé en D-LTAG en s'appuyant sur le *Penn Discourse TreeBank* (Webber, 2004). Ces deux solutions sont complémentaires.

5 Comparaison entre D-STAG et D-LTAG

D-STAG ressemble à D-LTAG - dans la version présentée dans (Forbes-Riley *et al.*, 2006) - dans la mesure où ces deux formalismes étendent une interface syntaxe/sémantique basée sur TAG au discours. Cependant, il existe une différence cruciale : D-LTAG n'utilise pas les relations de discours et ignore la distinction entre relations coordonnantes et subordonnantes. Les formes logiques des discours sont calculées par le même procédé que celui utilisé pour calculer les formes logiques des phrases. Ceci donne à D-LTAG une homogénéité certaine mais l'empêche de bénéficier des résultats apportés par les théories sur le discours.

Il existe une autre différence entre D-STAG et D-LTAG : les analyses syntaxiques des discours sont différentes car en D-STAG les connecteurs de discours ancrent des arbres élémentaires avec deux arguments, tandis qu'en D-LTAG ils peuvent ancrer des arbres avec un seul argument (qui est fourni *structurellement*, l'autre étant fourni *anaphoriquement* (Webber *et al.*, 2003; Webber, 2004)), e.g. l'arbre pour *ensuite* n'a qu'un seul argument[11].

[10]Dans un discours de la forme P_1. P_2. dans lequel les phrases P_1 et P_2 ne sont pas liées par un item lexical (un connecteur de discours), on suppose que P_2 comporte le connecteur vide ϵ, ce qui s'écrit P_1. ϵ P_2. Une façon différente mais équivalente de voir les choses consiste à considérer le point séparant P_1 et P_2 comme un connecteur de discours (Danlos, 1998).

[11]A titre d'illustration, l'analyse syntaxique de (1) fournie par D-LTAG est donnée ci-contre ;

6 Conclusion

D-STAG est un formalisme qui étend une interface syntaxe/sémantique développée en STAG au niveau du discours et qui prend en compte la pragmatique du discours. Une grammaire D-STAG génère des analyses qui peuvent être interprétées comme des structures RST ou SDRT. Les formes logiques pour les discours sont calculées de façon déterministe à partir des analyses compositionnelles (arbres de dérivation). Dans (Danlos, 2007), nous montrons comment traiter en D-STAG des phénomènes complexes de portée.

La recherche future concernera les discours dans lesquels un argument d'une relation de discours provient d'un segment de texte discontinu, ce qui arrive avec la relation Attribution quand un de ses arguments est enchâssé dans l'autre (qui est de ce fait discontinu). Nous pensons que la relation d'adjonction sera d'un grand secours pour ces cas. Ceux-ci demanderont d'entremêler les grammaires STAG phrastique et discursive, alors qu'elles ont été considérées comme fonctionnant séquentiellement (*en pipe-line*) dans cet article.

Références

ASHER N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht : Kluwer.

ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge : Cambridge University Press.

ASHER N. & VIEU L. (2005). Subordinating and coordinating discourse relations. *Lingua*, **115**(4), 591–610.

CARLSON L., MARCU D. & OKUROWSKI M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. VAN KUPPEVELT & R. SMITH, Eds., *Current Directions in Discourse and Dialogue*, p. 85–112. Kluwer Academic Publishers.

DANLOS L. (1998). G-TAG : un formalisme lexicalisé pour la génération de textes inspiré de TAG. *Revue TAL*, **39**(2).

DANLOS L. (2004). Sentences with two subordinate clauses : syntactic and semantic analyses, underspecified semantic representation. In *Proceedings of TAG+7*, p. 140–147, Vancouver.

DANLOS L. (2007). Flexible composition in D-STAG. In *Proceedings of MTT'07*, Klagenfurt, Austria.

FORBES-RILEY K., WEBBER B. & JOSHI A. (2006). Computing discourse semantics : The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, **23**(1).

JOSHI A. (1985). Tree-adjoining grammars. In D. DOWTY, L. KARTTUNEN & A. ZWICKY, Eds., *Natural language parsing*, p. 206–250. Cambridge University Press.

KAMP H. & REYLE U. (1993). *From Discourse to Logic*. Dordrecht : Kluwer Academic Publishers.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

MARCU D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

NESSON R. & SHIEBER S. (2006). Simpler TAG semantics through synchronization. In *Formal Grammars*, Malaga.

POLANYI L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, **12**, 601–638.

SHIEBER S. (1994). Restricting the weak-generative capacity of synchronous tree-adjoining grammars. *Computational Intelligence*, **10**(4), 371–385.

SHIEBER S. & SCHABES Y. (1990). Synchronous tree-adjoining grammars. In *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, p. 253–258, Helsinki.

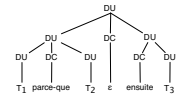
STEDE M. (2007). RST revisited : Disentangling nuclearity.

TABOADA M. & MANN W. (2006). Rhetorical structure theory : Looking back and moving ahead. *Discourse Studies*, **8**(3), 423–459.

WEBBER B. (2004). DTAG : extending lexicalized TAG to discourse. *Cognitive Science*, **28**(5), 751–779.

WEBBER B. L., JOSHI A., STONE M. & KNOTT A. (2003). Anaphora and discourse structure. *Computational Linguistics*, **29**(4), 545–587.

elle comporte trois nœuds étiquetés DC, un pour *parce que*, un pour ϵ, et un pour *ensuite*. Elle est différente de l'analyse syntaxique de (1) fournie par D-STAG - Figure 5(a) - qui ne comporte que deux nœuds étiquetés DC.



Session
Traduction & alignement

Collocation translation based on sentence alignment and parsing

Violeta SERETAN, Eric WEHRLI
Language Technology Laboratory (LATL) - University of Geneva,
2 Rue de Candolle, 1211 Geneva, Switzerland
{Violeta.Seretan,Eric.Wehrli}@lettres.unige.ch

Résumé. Bien que de nombreux efforts aient été déployés pour extraire des collocations à partir de corpus de textes, seule une minorité de travaux se préoccupent aussi de rendre le résultat de l'extraction prêt à être utilisé dans les applications TAL qui pourraient en bénéficier, telles que la traduction automatique. Cet article décrit une méthode précise d'identification de la traduction des collocations dans un corpus parallèle, qui présente les avantages suivants : elle peut traiter des collocation flexibles (et pas seulement figées) ; elle a besoin de ressources limitées et d'un pouvoir de calcul raisonnable (pas d'alignement complet, pas d'entraînement) ; elle peut être appliquée à plusieurs paires des langues et fonctionne même en l'absence de dictionnaires bilingues. La méthode est basée sur l'information syntaxique provenant du parseur multilingue Fips. L'évaluation effectuée sur 4000 collocations de type verbe-objet correspondant à plusieurs paires de langues a montré une précision moyenne de 89.8% et une couverture satisfaisante (70.9%). Ces résultats sont supérieurs à ceux enregistrés dans l'évaluation d'autres méthodes de traduction de collocations.

Abstract. To date, substantial efforts have been devoted to the extraction of collocations from text corpora. However, only a few works deal with the subsequent processing of results in order for these to be successfully integrated into the NLP applications that could benefit from them (e.g., machine translation). This paper presents an accurate method for identifying translation equivalents of collocations in parallel text, whose main strengths are that : it can handle flexible (not only rigid) collocations ; it only requires limited resources and computation (no full alignment, no training needed) ; it deals with several language pairs, and it can even work when no bilingual dictionary is available. The method relies heavily on syntactic information provided by the Fips multilingual parser. Evaluation performed on 4000 verb-object collocations for different language pairs showed an average accuracy of 89.8% and a reasonable coverage (70.9%). These figures are higher that those reported in the evaluation of related work in collocation translation.

Mots-clés : traduction de collocations, extraction de collocations, parsing, alignement de textes.

Keywords: collocation translation, collocation extraction, parsing, text alignment.

1 Introduction

Collocations constitute a subclass of phraseological units (or *multi-word expressions*) that received particular attention in several research fields—e.g., second language learning, lexicography and natural language processing—both because of their massive presence in language and because of their specific features : although they look similar to regular constructions, they are unpredictable for non-native speakers and usually do not have a literal translation. Consider, for instance, the collocation *to break a record*. A non-native speaker of English would hardly choose *break* as the support verb for the noun *record*. Moreover, this collocation does not translate in a word-for-word fashion into French (**casser un record*), but as *battre un record* (lit., *to beat a record*).

For several decades already, sustained efforts have been put into developing methods for the automatic extraction of collocations from text corpora, as well as into the evaluation of extraction results ; see (Seretan & Wehrli, 2006) for a thorough review. Considerably less work deals instead with the post-processing of extracted collocations and with their further integration into other NLP applications, like machine translation, natural language generation, parsing, word sense disambiguation, or text classification. Among the few exceptions, we can mention works dealing with the semantic classification of collocations (Wanner *et al.*, 2006), the extraction of synonymous collocations (Wu & Zhou, 2003), the translation of collocations (Smadja *et al.*, 1996; Lü & Zhou, 2004), or the use of collocations in machine translation (Orliac & Dillinger, 2003), in natural language generation (Heid & Raab, 1989), and in text classification (Williams, 2002). Unfortunately, these efforts remained generally isolated and at an early stage of development, despite the largely acknowledged critical role played by such expressions in many NLP tasks (Sag *et al.*, 2002) and the continuous improvement of extraction techniques.

This paper describes a method for obtaining collocation equivalents from sentence-aligned texts that is based on the parsing of source and target sentences. The main advantages of this method are that it can deal with flexible (as opposed to rigid) collocations ; it does not require an extensive computation or huge training resources ; and it does not rely crucially on the availability of bilingual dictionaries.

The paper is organized as follows. Section 2 presents a review of previous work on collocation translation. Section 3 introduces our method and briefly describes the processing modules on which it relies (the multilingual parser, the collocation extractor, and the sentence aligner). Experimental results, an evaluation of the method and the error analysis are presented in section 4. Section 5 concludes the paper by discussing the relative merits of the newly introduced method with respect to existing methods, and by pointing out the ways in which this method can be improved in order to attain better performance.

2 Previous Collocation Translation Work

Corpus-based collocation translation has previously been dealt with in a number of works. One of the earliest is (Kupiec, 1993), that identifies noun phrase correspondences between English and French from Hansard parallel corpus. Both source and target corpora are POS-tagged, then NPs are detected with a finite-state recognizer. For mapping correspondences, the author uses Expectation Maximization (EM), an iterative re-estimation algorithm. The method was evaluated on a small set of 100 NPs, and achieved 90% accuracy. Also, Van der Eijk (1993) performed

a similar experiment for Dutch to English, but the reported accuracy was lower, since the evaluation was performed on a larger test set. This method differs from (Kupiec, 1993) in that it uses relative frequencies for computing the mappings between source and target terms.

Pursuing the same goal, Dagan and Church (1994) use a word aligner to propose (multiple) candidate translations for rigid noun phrases. Unlike the previous approaches, their system, TERMIGHT, has the advantage of being able to find translations even for infrequent terms. But like the preceding systems, it deals with rigid constructions only.

Later, the first proper collocation translation system, Champollion, has been implemented by Smadja *et al.* (1996). It relies on Xtract (Smadja, 1993) for detecting source collocations in English, then it applies a statistical correlation metric, namely the Dice coefficient, for identifying their translation equivalents in the aligned French sentences in Hansard corpus. Noticeably, this system can also deal with flexible collocations (e.g., verbal phrases). It requires an additional post-processing step in which the order of words in a flexible collocation is decided, as no syntactic information is available. The system has been evaluated by three human annotators, and showed a precision of 77% and, respectively, 61% on two different test sets of 300 collocations each.

Finally, the work of Lü and Zhou (2004) can deal with flexible collocations as well ; moreover, these are validated syntactic constituents, since extracted with a dependency parser. The syntactic types considered are verb-object, adjective-noun, and adverb-verb. Collocations are extracted from monolingual corpora in English and Chinese by applying the log-likelihood ratios statistical test on the dependency pairs identified. The translation is then performed with a statistical translation model that estimates word translations with EM. The head and the dependent word are assigned uneven probabilities, while the dependency relation is considered to be preserved across languages. The method (whose reported coverage is 83.98%) has been evaluated on a test set of 1000 randomly selected collocations. It achieved between 50.85% and 68.15% accuracy, depending on the syntactic type.

3 Translating Collocations Using Parsing Information

3.1 The method

The translation procedure we developed involves a series of steps relying on other processing modules, shortly described below. The procedure assumes that a parallel corpus is available, and that both the source and target languages are supported by the parser. First, collocations are extracted from the source corpus by using a hybrid extraction procedure (section 3.3) that combines a standard statistical technique with the deep syntactic analysis performed with the Fips parser (section 3.2). In the next step, for each collocation pair extracted, a limited number of sentence contexts is selected amongst all its contexts of occurrence in the source corpus ; in our present experiments, we considered a maximum of 50 contexts for each collocation.

The source sentences are then aligned using a sentence-aligner (section 3.4) and the equivalent target sentences are gathered into a small corpus specific to each source collocation. As the whole translation procedure is automatic, no manual validation is performed on the resulting sentence alignments. The corpus of target sentences is subsequently processed with Fips, and candidate collocation pairs are extracted using the same method as in the case of source collo-

cations. Finally, to find out the translation of the source collocation given these candidate pairs, we apply a matching procedure, which is described in section 3.5.

3.2 The Fips parser

Fips is a deep symbolic parser based on generative grammar concepts that was developed at LATL over the last decade (Wehrli, 2006). It is written in Component Pascal and adopts an object-oriented implementation design allowing to couple language-specific processing with a generic core module. The parsing algorithm proceeds in a bottom-up fashion, by applying (general or language-specific) licensing rules, by treating alternatives in parallel, and by using pruning heuristics. The parser currently supports the following languages: English, French, Spanish, Italian, German (other languages are under development as well).

In Fips, each syntactic constituent is represented as a simplified X-bar structure of the form $[_{XP}$ L X R] with no intermediate level, where X is a variable ranging over the set of lexical categories[1]. L and R stand for (possibly empty) lists of, respectively, left and right subconstituents. The lexical level contains detailed morphosyntactic and semantic information available from the manually-built lexicons. In the structures returned by the parser, extraposed elements (interrogative phrases, relative pronouns, clitics, etc.) are coindexed with empty constituents in canonical positions (i.e., typical argument or adjunct positions).

3.3 Collocation extraction with Fips

The first step in the collocation extraction process is the identification of collocation candidates. Once a sentence has been parsed by Fips, the resulting structure is checked for potential collocational pairs, by recursively examining all the pairs consisting of the head of a phrase and an element of one of its left or right subconstituents. Those pairs that satisfy certain constraints are retained as valid collocation candidates. The constraints may refer both to the lexical items individually, and to the combination as a whole. For instance, proper nouns and auxiliary verbs are ruled out, and combinations are considered valid if in a configuration like the following[2]: A-N: *wide range*, N-A: *work concerned*, N-N: *food chain*, N(subject)-V: *rule applies*, V-N(object): *strike balance*, V-P: *reflect upon*, V-P-N(argument or adjunct): *comply with rule*, N-P-N: *fight against terrorism*, V-A: *steer clear*, V-Adv: *desperately need*, Adv-A: *highly controversial*, A-P: *favourable to*, coordinated A-A: *nice and warm*, coordinated N-N: *part and parcel*. It is worth noting that each lexical item may in turn be a complex lexeme (e.g., a compound or a collocation), like *death penalty* in *abolish the death penalty* ; such a lexeme can be recognized as a single lexical item by Fips, if is is present in its lexicon.

In the second extraction step, the candidate pairs that have been identified in step one are partitioned into syntactically-homogeneous classes, then log-likelihood ratios test (Dunning, 1993) is applied on each class. *Log-likelihood ratios* (LLR) is a statistical hypothesis test that can be used to identify statistically-significant pairs among candidates (i.e., collocations) based on lexical co-occurrence evidence organized in a so-called contingency table, for each two lexical items making up a candidate pair. This table lists, essentially, the joint frequency of the two

[1]The lexical categories are N(oun), Adj(ective), V(erb), P(reposition), Adv(erb), C(onjonction), Inter(jection), to which we add the two functional categories T(ense) and F(unctional).

[2]The list of configurations is not exhaustive. It is, in fact, growing as more corpus evidence is considered.

items, the marginal frequency of each item, and the total number of pairs in the corresponding class. The result of extraction is represented by the initial list of candidate pairs ranked according to the LLR score ; the higher the score, the more likely that the pair constitutes a collocation. More details about the extraction procedure can be found in (Nerima *et al.*, 2003).

3.4 Sentence alignment

Given a parallel corpus, a sentence alignment tool finds, for each source sentence, the corresponding sentence in the target corpus (i.e., the translation equivalent of that sentence, or the target sentence). State-of-the-art sentence aligners are based on the char-length of words in sentences, on lexical clues (e.g., numbers, cognate words) and possibly exploit the macro-structure of documents (titles, sections, paragraphs)[3].

We employed our own sentence-aligner based on lexical clues and on context-size matching for paragraph detection, followed by a one-by-one sentence alignment within the aligned paragraphs. The method, which is fully described in (Nerima *et al.*, 2003), has the advantage of computing a partial, on-the-fly sentence alignment for a given source sentence identified by the file position of a word inside that sentence. This aligner is best suited for our purpose, as it allows the rapid identification of the target sentence given an item of the source collocation, without us being forced to align the whole source and target documents. Although the aligner's accuracy is not perfect (between 88% and 93.5%), the translation results obtained with our procedure suggest that it is nonetheless satisfactory for this specific task.

3.5 The matching procedure

To actually translate a collocation, we try to match it against the collocation candidates extracted from the associated target sentences. Like in (Lü & Zhou, 2004), we assume that the mapping between the source collocation and the target collocation preserves the syntactical relation involved, meaning, for instance, that a verb-object collocation in French translates into a verb-object collocation in English[4]. Therefore, we first perform a syntactic filter on the target candidate pairs that retains only the appropriate pairs, i.e., those that involve the same syntactic relation as the source collocation.

We then (optionally) apply a 'semantic' filter as follows: first, we derive from the syntactic type information about the semantic head of a collocation (usually called *base*). For instance, the base of a verb-object collocation is the object, that of an N-A collocation is the noun N, etc. Collocations are known to preserve the translation of the base, while the translation of collocate can vary across languages. For instance, in translating *break a record* into French, the noun *record* is preserved, while the verbal collocate *break* (*casser*) is transformed into *battre*. Whenever translation information for the base of the source collocation is available in our bilingual dictionaries, we consider all its possible translations and we apply a filter on the target candidate pairs accordingly ; otherwise, this filter is skipped. Finally, we select as target collocation (i.e., as translation of the source collocation) the most frequent pair among the remaining pairs, after the filters described above have been applied.

[3]Lack of space prevents us from providing more details here.

[4]This is obviously not always the case. Yet, this (simplifying) assumption was shown by Lü and Zhou (2004) to hold in the majority of cases.

4 Results and Evaluation

4.1 Translation experiment

The translation experiment described in this paper was performed on collocations extracted from a parallel corpus in four languages (English, French, Italian and Spanish), which is a subpart of Europarl parallel corpus of European Parliament proceedings (Koehn, 2005). It contains 62 files in each of the four languages that correspond to the complete proceedings for the year 2001. Table 1 displays several statistics on the corpus (rows 1–4) and on the collocations extracted with the method presented in section 3.3 (rows 5–6).

Row	Statistics	English	French	Italian	Spanish
1	Size (MB)	21.4	23.7	22.9	22.7
2	Tokens	4158622	4770835	4134549	4307360
3	Sentences	161802	162671	160906	172121
4	Average sentence length (tokens)	25.7	29.3	25.7	25
5	Total collocation pairs extracted	851500	988918	880608	901224
6	Distinct collocation pairs extracted	333428	327366	333848	315532
7	V-O pairs in translation set (500 distinct)	28005	27058	25787	23003
8	Frequency range for pairs in translation set	5–852	6–784	7–1085	6–480

TAB. 1 – Experimental statistics: corpus size, collocations extracted, translation sets size.

From the whole set of collocations extracted, we have chosen for our translation experiment the top 500 verb-object collocations obtained in each language. These 500 collocation types correspond to many more instances occurring in the corpus ; row 7 of Table 1 displays the total number of instances in each translation set, and row 8 shows the frequency range for the collocation types in each set.

The translation method described in section 3 has been applied on these translation sets in each of the possible directions. Therefore, for the 4 languages considered, there are 12 language pairs on which the method was applied. Several (randomly chosen) translations obtained are listed in Table 2.

4.2 Evaluation of results and error analysis

The random examples of translations shown in Table 2 suggest that the accuracy of our method is quite high. In fact, the evaluation performed until now shows that surprisingly good results can be achieved with this rather simple method.

The results obtained for a couple of language pairs in the translation experiment presented here have been thoroughly checked by a human judge. Whenever necessary, the contexts of the source collocation in the original documents have been inspected and confronted against the target sentences with the help of a concordancer connected to our collocation extractor and sentence aligner (Seretan *et al.*, 2004). The accuracy results for the language pairs evaluated until now are shown in the second column of Table 3. As it can be seen, comparable accuracy is achieved for the language pairs for which a bilingual (mono-lexeme) dictionary is available (90.9% to 94.1%). When such a dictionary is not available, results are worse, but still satisfactory (82.4% to 85.8%). The average accuracy obtained is 89.8%.

	Source collocation	Translation	Source collocation	Translation
En-Fr	express satisfaction	exprimer satisfaction	accroître transparence	increase transparency
	create condition	créer condition	corriger erreur	*make mistake
	improve safety	améliorer sécurité	perdre vie	lose life
	transpose directive	transposer directive	devenir réalité	become reality
	draw conclusion	tirer conclusion	remercier collègue	thank colleague
En-It	ask question	porre domanda	soddisfare requisito	meet requirements
	have opportunity	avere occasione	modificare direttiva	amend directive
	vote reason	*votare relazione	creare situazione	create situation
	thank presidency	ringraziare presidenza	apportare contributo	make contribution
	congratulate Mrs.	*svolgere lavoro	garantire livello	ensure level
En-Es	achieve goal	alcanzar objetivo	ser placer	be pleasure
	address question	abordar cuestión	recibir respuesta	receive reply
	draw list	hacer lista	ocupar lugar	take place
	play role	desempeñar papel	suspender sesión	suspend sitting
	find way	encontrar salida	*sobrar base	*draw inspiration
Fr-It	déployer effort	compiere sforzo	avere compito	avoir tâche
	transposer directive	recepire direttiva	commettere reato	commettre délit
	demander parole	chiedere parola	approvare risoluzione	adopter résolution
	vacciner animal	vaccinare animale	prendere impegno	prendre engagement
	ménager effort	lesinare sforzo	effettuare studio	mener étude
Fr-Es	poursuivre effort	continuar esfuerzo	emitir dictamen	donner avis
	éradiquer terrorisme	erradicar terrorismo	examinar cuestión	examiner question
	produire électricité	generar electricidad	hacer distinción	faire distinction
	jouer rôle	desempeñar papel	marcar hito	représenter étape
	lever obstacle	eliminar obstáculo	traspasar frontera	passer frontière
It-Es	rispettare principio	respetar principio	promover desarrollo	promuovere sviluppo
	avere impressione	tener impresión	manifestar gratitud	*ringraziare relatore
	approvare posizione	aprobar posición	tener intención	avere intenzione
	rispettare impegno	respetar compromiso	acumular retraso	accumulare ritardo
	affrontare problema	abordar problema	hacer observación	fare osservazione

TAB. 2 – Randomly chosen translation results (incorrect translations or invalid source collocations are marked with an asterisk).

The third column in Table 3 shows the method's coverage. This corresponds, in our case, to the ratio of collocation pairs for which a translation was proposed (70.9% on average). Our method does not propose a translation for a collocation when there are several translation candidates with the same frequency (previous examination of results indicated that taking all candidates in a tie introduces more noise than good translation alternatives), or when there are no candidates left after the two filters have been applied. This situation might occur for the lower frequency collocations ; our translation sets contain collocations whose frequency is as low as 5–7.

Table 3 also reports the impact of frequency on our method's performance. Accuracy and coverage have been computed separately for three frequency intervals (we distinguished between high-, medium-, and low-frequency data, corresponding to the following frequency ranges: 31–50, 16–30, and 1–15). The results obtained suggest that only a minor decrease in accuracy is observed as the frequency decreases, while the coverage is more drastically affected.

Error analysis performed on the evaluated collocations highlighted a series of issues that affect the performance of our method. Since we apply no restriction on the collocate other than the

Language pair	Accuracy				Coverage				Dictionary
	All	31–50	16–30	1–15	All	31–50	16–30	1–15	
English-French	94.1	95.6	93.3	89.8	71.4	75.8	70.7	58.3	+
English-Italian	85.8	86.2	89.3	75.7	64.8	75.5	57.1	44.0	-
French-English	92.8	94.7	89.3	92.7	72.2	80.0	65.5	59.4	+
French-Italian	92.8	91.8	96.5	87.8	72.2	79.6	66.1	59.4	+
French-Spanish	90.9	92.0	90.9	85.7	75.0	81.5	70.8	60.9	+
Italian-English	82.4	87.6	75.2	74.1	63.6	72.9	58.3	41.5	-
Italian-French	94.1	97.0	88.9	93.1	67.8	79.2	60.0	44.6	+
Italian-Spanish	85.3	89.5	80.0	77.8	80.0	89.8	75.0	55.4	-
Average	89.8	91.8	87.9	84.6	70.9	79.3	65.4	53.0	

TAB. 3 – Evaluation results (for the whole translation sets and for different frequency intervals).

syntactic filter[5], our method could propose a wrong candidate if this happens to occur systematically in the context of the right collocate and has the same syntactic type. For instance, a sentence like the one in example 1 below occurs a lot in the corpus. Our method proposes an incorrect translation for *reprendre séance*, namely **suspend a sitting*, since *suspend a sitting* occurs systematically in the context of the right candidate *resume a sitting*, and it has the same syntactic type, verb-object. Moreover, it is easier to analyse than the right candidate, which is in turn more susceptible to be missed by the parser.

1. *The sitting was suspended at 1 p.m. and resumed at 3 p.m.*
2. *This compromise formula broke the deadlock in Council and opened the door to the approval of the negotiating directives.*
   Tale formula di compromesso riuscì a sbloccare la situazione di stallo nel Consiglio e spianò la strada all'approvazione delle direttive negoziali.
3. *En tant qu'homme de science, je voudrais faire une autre remarque, Monsieur le Commissaire.*
   As a scientist, I would like to make another point, dear Commissioner.

A more recurrent situation is that in which one of the items in a collocation is lexicalized across languages, or the whole collocation is lexicalized, i.e., paraphrased as a single word. Example 2 shows the item *situazione di stallo* in the target collocation *sbloccare la situazione di stallo*, which in English is lexicalized as the single word *deadlock*. Our method incorrectly translates *break the deadlock* into *sbloccare la situazione* instead of *sbloccare la situazione di stallo*[6], since the parser does not recognize *situazione di stallo* as a lexical unit. Once this unit is added in the parser's lexicon, our method could find the good translation. An example in which the whole source collocation is lexicalized is *manifestar gratitud* shown in Table 2, whose Italian equivalent is a single word, *ringraziare*. Our method cannot handle such situations, and wrongly adds an object (**ringraziare relatore*) to the otherwise good verbal translation identified.

Quite frequent are also the situations in which the translation of a collocation is difficult to find due to the free human translations the parallel corpus contains: one can find too vague paraphrases: *hold any necessary debates*/ *participer à tous les débats nécessaires* ; omissions of a collocation item: *is to hold a debate*/ *avec le débat* ; complete omission of the collocation: *Once we have held the debate*/ *À ce moment-là*, etc. Similar problems are posed by the syntactic

[5]That is, we do not apply a semantic filter (as in the case of the base word) or finer syntactic constraints, such as imposing a syntactic structure matching between the source and target contexts.

[6]Although this kind of translations can be interpreted as partly correct, we marked them as incorrect as we did not use a gradual scale in our evaluation.

structure changes across languages (e.g., V-N vs. V-P-N: *attend meeting/ participer à reunion*, V-N vs. V-A: *pay attention/ être attentif*), or, interestingly, by negation: *It is no easy task/ Il s'agit d'une rude tâche*.

Clearly, the parsing and alignment errors as well as the coverage of the bilingual lexicons also affect our method's accuracy. If parsing and alignment errors do not influence much (as long as they can be compensated by looking at other contexts[7]), dictionary coverage problems have more drastic consequences: if a translation for a base word is missing from the dictionary and the corpus systematically contains exactly that translation, the method cannot propose a translation for the source collocation. For instance, our French-English dictionary lists, for the entry *remarque*, the following translations: *remark, comment, note*. However, the translation *point* is also needed in order for our method to identify the translation of *faire remarque* from contexts like in example 3 above, that involves the pair *make point*.

5 Conclusion

Thanks to the methodology used, the method we presented has several advantages over existing collocation translation techniques. Unlike (Kupiec, 1993; van der Eijk, 1993; Dagan & Church, 1994), it can handle flexible collocations. Unlike (Smadja *et al.*, 1996), it does not require the postprocessing of results (lexical re-ordering), since target collocations are extracted with a parser. With respect to (Lü & Zhou, 2004), it deals with more syntactic types and more languages ; it does not depend crucially on a bilingual dictionary ; it only uses mono-lexeme translations for the base word (since most of the times the collocate cannot be translated literally) ; and it is considerably simpler. In addition, it only requires several sentence contexts for a collocation, as opposed to the huge textual resources and the expensive training required by state-of-the-art phrase aligners developed in relation with statistical translation[8].

A limitation of our method is that it relies on a parallel corpus ; on the contrary, (Lü & Zhou, 2004) does not. However, in this setting our method was shown to produce quite accurate results, which suggest that adding parsing information is at least as helpful as using sophisticated statistical techniques. The method can be improved by defining syntactic configuration mappings between languages (in order to account for structure changes across languages, as those mentioned in section 4.2), by increasing the dictionaries coverage, and by including multi-word units in the parser's lexicon. Furthermore, its evaluation must be extended to other syntactic types, preferably once the syntactic mappings will be defined.

Acknowledgements

Part of the research described in this paper has been supported by a grant from the Swiss National Science Foundation (No. 101412-103999).

[7] We measured the performance of our method on low-frequency data in a separate evaluation experiment, and found that the accuracy is still acceptable for collocations occurring exactly 10 and 5 times in the corpus (85.7% and, respectively, 72.0%), but it drops to 39.1% when frequency is equal to 3 (a number of 100 English-French translation pairs have been investigated for each frequency level). The coverage is drastically affected (42%, 25% and 23%). One important reason for this degradation are the unsystematic translations found in the parallel corpus.

[8] Note that the phrase translations produced in PBSMT (Phrase-Based Statistical Machine Translation) do not have a linguistic interpretation/motivation, since a phrase simply means there any sequence of words.

Références

DAGAN I. & CHURCH K. (1994). TERMIGHT: Identifying and translating technical terminology. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, Stuttgart, Germany.

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

HEID U. & RAAB S. (1989). Collocations in multilingual generation. In *European Chapter of the Association for Computational Linguistics (EACL'89)*, p. 130–136, Manchester, England.

KOEHN P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of The Tenth Machine Translation Summit (MT Summit X)*, p. 79–86, Phuket, Thailand.

KUPIEC J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *31st Annual Meeting of the Association for Computational Linguistics*, p. 17–22, Columbus, Ohio, U.S.A.

LÜ Y. & ZHOU M. (2004). Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, p. 167–174, Barcelona, Spain.

NERIMA L., SERETAN V. & WEHRLI E. (2003). Creating a multilingual collocation dictionary from large text corpora. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, p. 131–134, Budapest, Hungary.

ORLIAC B. & DILLINGER M. (2003). Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX*, p. 292–298, New Orleans, Lousiana, U.S.A.

SAG I. A., BALDWIN T., BOND F., COPESTAKE A. & FLICKINGER D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, p. 1–15, Mexico City.

SERETAN V., NERIMA L. & WEHRLI E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora. In *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*, p. 755–766, Lorient, France.

SERETAN V. & WEHRLI E. (2006). Multilingual collocation extraction: Issues and solutions. In *Proceedings of COLING/ACL Workshop on Multilingual Language Resources and Interoperability*, p. 40–49, Sydney, Australia. 2006.

SMADJA F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, **19**(1), 143–177.

SMADJA F., MCKEOWN K. & HATZIVASSILOGLOU V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, **22**(1), 1–38.

VAN DER EIJK P. (1993). Automating the acquisition of bilingual terminology. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, p. 113–119, Utrecht, The Netherlands.

WANNER L., BOHNET B. & GIERETH M. (2006). Making sense of collocations. *Computer Speech & Language*, **20**(4), 609–624.

WEHRLI E. (2006). TwicPen: Hand-held scanner and translation software for non-native readers. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, p. 61–64, Sydney, Australia: Association for Computational Linguistics.

WILLIAMS G. (2002). In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, **7**(1), 43–64.

WU H. & ZHOU M. (2003). Synonymous collocation extraction using translation information. In *Proceeding of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, p. 120–127, Sapporo, Japan.

Utilisation d'une approche basée sur la recherche cross-lingue d'information pour l'alignement de phrases à partir de textes bilingues Arabe-Français

Nasredine SEMMAR, Christian FLUHR
CEA, LIST, LIC2M
18 route du Panorama, BP6, FONTENAY AUX ROSES, F- 92265 France
{nasredine.semmar,christian.fluhr}@cea.fr

Résumé. L'alignement de phrases à partir de textes bilingues consiste à reconnaître les phrases qui sont traductions les unes des autres. Cet article présente une nouvelle approche pour aligner les phrases d'un corpus parallèle. Cette approche est basée sur la recherche cross-lingue d'information et consiste à construire une base de données des phrases du texte cible et considérer chaque phrase du texte source comme une requête à cette base. La recherche cross-lingue utilise un analyseur linguistique et un moteur de recherche. L'analyseur linguistique traite aussi bien les documents à indexer que les requêtes et produit un ensemble de lemmes normalisés, un ensemble d'entités nommées et un ensemble de mots composés avec leurs étiquettes morpho-syntaxiques. Le moteur de recherche construit les fichiers inversés des documents en se basant sur leur analyse linguistique et retrouve les documents pertinents à partir de leur indexes. L'aligneur de phrases a été évalué sur un corpus parallèle Arabe-Français et les résultats obtenus montrent que 97% des phrases ont été correctement alignées.

Abstract. Sentence alignment consists in identifying correspondences between sentences in one language and sentences in the other language. This paper describes a new approach to aligning sentences from a parallel corpora. This approach is based on cross-language information retrieval and consists in building a database of sentences of the target text and considering each sentence of the source text as a query to that database. Cross-language information retrieval uses a linguistic analyzer and a search engine. The linguistic analyzer processes both documents to be indexed and queries to produce a set of normalized lemmas, a set of named entities and a set of nominal compounds with their morpho-syntactic tags. The search engine builds the inverted files of the documents on the basis of their linguistic analysis and retrieves the relevant documents from the indexes. An evaluation of the sentence aligner was performed based on a Arabic to French parallel corpus and results show that 97% of sentences were correctly aligned.

Mots-clés : alignement de phrases, corpus parallèle, recherche cross-lingue d'information.
Keywords: sentence alignment, parallel corpora, cross-lingual information retrieval.

1 Introduction

L'alignement de textes bilingues dont l'un est une traduction de l'autre consiste à mettre en relation des unités linguistiques ou logiques qui se correspondent dans les deux textes. Ces unités peuvent être des paragraphes, des phrases, des syntagmes, des mots, etc. L'alignement de textes permet l'élaboration de lexiques et de bases de données phraséologiques multilingues nécessaires pour la traduction et la terminologie. Plusieurs techniques d'alignement de textes ont été proposées (Gale, Church, 1991) (Brown et al., 1991) (Debili, Samouda, 1992) (Gaussier, 1995) (Melamed, 1996) (Fluhr et al., 2000).

Dans cet article, nous présentons un aligneur de phrases à partir de corpus parallèles utilisant une approche basée sur la recherche d'information cross-lingue et combinant plusieurs sources d'information (dictionnaire bilingue, longueurs des phrases, numéros d'ordre des phrases dans le corpus parallèle). Cet aligneur a été développé initialement pour aligner les corpus parallèles Français-Anglais, il a été ensuite adapté pour aligner les corpus des couples de langues Arabe-Français et Arabe-Anglais.

Nous présentons dans la section 2 les principaux composants du moteur de recherche cross-lingue du LIC2M, en particulier, nous nous focalisons sur les modules de l'analyse linguistique. Dans la section 3, nous décrivons le prototype de notre aligneur de phrases. Nous discutons dans la section 4 les résultats obtenus en alignant le corpus MD (Monde Diplomatique) de la campagne ARCADE II. La section 5 conclut notre étude et présente nos travaux futurs.

2 Le moteur de recherche cross-lingue

Le moteur de recherche cross-lingue permet, à partir d'une requête en une seule langue, de fournir des réponses trouvées dans des documents qui sont dans d'autres langues. Le moteur de recherche cross-lingue du LIC2M est composé d'un analyseur linguistique, d'un analyseur statistique, d'un reformulateur et d'un comparateur (Semmar et al., 2005).

2.1 L'analyse linguistique

L'analyse linguistique des documents et de la requête est un composant important dans le système de recherche d'information cross-lingue du LIC2M. L'analyseur linguistique LIMA (LIc2m Multilingual Analyser) est composé d'un ensemble de modules dont le nombre et la nature varient selon la langue traitée et un ensemble de ressources linguistiques. Selon que l'on traite l'arabe, le français ou le chinois, le système sait modifier le traitement et utiliser les ressources adaptées. Nous présentons dans les sections suivantes les modules et les ressources utilisés dans l'analyseur linguistique LIMA en se focalisant sur les traitements spécifiques à la langue arabe (Grefenstette et al., 2005).

2.1.1 Modules de traitement linguistique

Certains des ces modules sont utilisés pour le traitement de la majorité des langues traitées par LIMA. D'autres, plus spécifiques, ne sont utilisés que pour certaines langues.

1. La tokenisation qui découpe le texte en mots (tokens).

2. La consultation du dictionnaire des formes qui permet éventuellement de récupérer des informations linguistiques concernant les mots à reconnaître. Pour ceux qui sont semi voyellés ou non voyellés, cette consultation du lexique permet de récupérer les formes voyellées correspondantes, c'est à dire leurs alternatives orthographiques lorsqu'elles existent. Dans le cas par exemple du mot non voyellé مدرسة la recherche dans le dictionnaire donne les deux alternatives orthographiques suivantes: مَدْرَسَة "Ecole" (Nom commun féminin singulier) et مُدَرِّسَة "Institutrice" (Nom commun féminin singulier).

3. Lorsque leur forme de surface le permet, les mots sont segmentés en proclitique-radical-enclitique ou en proclitique-radical ou en radical-enclitique ou en proclitique-enclitique. Ce module de segmentation n'est utilisé que pour l'Arabe et l'Espagnol. Par exemple, le mot مدرسه est candidat au découpage مُدَرِّس + ه "Instituteur" + "lui, à lui".

4. Les expressions idiomatiques sont ensuite reconnues et regroupées pour être considérées comme un seul mot dans le graphe d'analyse. Cette reconnaissance se fait à l'aide de règles de déclencheurs qui sont généralement des lemmes. Ces règles permettent par exemple de reconnaître les noms de mois arabes ذو العِقْدَة et جُمَادَى الأولى comme des mots uniques.

5. Si, après ces étapes, un mot reste inconnu, le système lui attribue une/des catégorie(s) par défaut en s'appuyant généralement sur des informations révélées par sa forme de surface.

6. Après cette analyse morphologique, la majorité des mots restent ambigus notamment à cause du nombre élevé des voyellations possibles. Le rôle du désambiguïseur morpho-syntaxique est ensuite de réduire le nombre des ambiguïtés en utilisant des matrices de désambiguïsation. Ce sont des matrices de bi-grammes et tri-grammes obtenues à partir d'un corpus de 13 200 mots pour l'Arabe et de 25 000 mots pour le Français. Ce corpus est étiqueté et désambiguïsé manuellement. La précision du désambiguïseur morpho-syntaxique est d'environ 91% pour l'Arabe et de 94% pour le Français.

7. Une analyse syntaxique tente ensuite, par un jeu de règles écrites à la main, d'établir les relations de dépendance entre les mots dans un même syntagme et entre les syntagmes dans une même phrase. Par exemple, dans la chaîne nominale توزيع المياه الصالحة "Distribution des eaux potables", l'analyse syntaxique considère que cette chaîne est un mot composé des mots توزيع "Distribution" (nom commun), مياه "Eaux" (nom commun) et صالحة "Potables" (adjectif).

8. Une reconnaissance des entités nommées est ensuite activée. Cette étape de l'analyse utilise des fichiers de listes ainsi que de règles de déclencheurs pour reconnaître des entités telles que les noms de personnes, d'organisations, de produits et de lieux, les dates ainsi que les unités de mesure. Ainsi, un énoncé comme الأوّل مِن شَهْر مَارس "Le premier du mois de Mars" est reconnu comme une date et الشّرْق الأوْسَط "Le Moyen-Orient" est reconnu comme un nom de lieu.

2.1.2 Ressources linguistiques

Pour le traitement de l'arabe, le système dispose de ressources lexicales et grammaticales suivantes:

- Un dictionnaire de formes qui contient toutes les formes fléchies et dérivées simples des mots en arabe. Ce dictionnaire est obtenu par un fléchisseur automatique développé au sein du LIC2M (Debili, Zouari, 1985). Cet outil produit 3 164 000 entrées à partir de 14 000 lemmes (noms, adjectives et verbes). Le dictionnaire final contient également les listes fermées comme les pronoms, les prépositions, les nombres, etc. Les mots du dictionnaire ont deux sortes d'entrées: des formes complètement voyellées ou complètement dévoyellées. Seules les entrées voyellées possèdent des informations linguistiques (catégorie, genre, nombre, etc.). Les entrées non voyellées, qui sont ambiguës par nature, ne possèdent que des pointeurs vers les entrées voyellées correspondantes et donc leur informations linguistiques comme il a été montré dans l'exemple de مدرسة plus haut. Le contenu du dictionnaire ne permet pas seulement d'attribuer des voyellations aux mots non voyellés mais aussi de proposer les différentes alternatives concernant certaines lettres comme c'est la cas pour آ إ أ qui sont trois manières différentes d'écrire la lettre ا et pour ىء et ءو qui sont deux alternatives à l'écriture pour les lettres ئ et ئ.

- Un dictionnaire de proclitiques ainsi qu'un dictionnaire d'enclitiques simples et composés. La même structure est attribuée à ces entrées, c'est à dire une forme voyellée et une forme non voyellée correspondante. Par exemple, le proclitique أفبِ est décomposé en trois parties أ +فَ + بِ.

- Des dictionnaires bilingues pour toutes les paires de langues traitées par le système sont également disponibles. Ces dictionnaires permettent la reformulation bilingue dans le cadre de la recherche d'information cross-lingue. Le dictionnaire bilingue Arabe-Français est composé de 120 000 entrées validées manuellement.

2.2 Analyse statistique

L'analyse statistique consiste à attribuer un poids aux mots simples et aux mots composés sur l'ensemble des documents indexés, selon le "degré d'information" qu'ils contiennent. Ce poids est lié à l'hétérogénéité de répartition du terme dans la base de documents. Il sera maximum si le terme est complètement discriminant, c'est-à-dire s'il apparaît dans un seul document, et minimum s'il n'est pas discriminant et apparaît dans tous les documents (Andreewsky et al., 1981).

2.3 Reformulation de la requête

Dans certains cas, l'analyse linguistique et l'analyse statistique expliquées ci-dessus ne suffisent pas à établir un lien entre la requête et les documents pertinents. Dans ce cas, il est nécessaire d'ajouter un élément sémantique au processus sur la base de la requête originale afin d'inférer ce que recherche l'utilisateur. Il s'agit donc d'étendre la requête posée en utilisant d'autres formulations de l'idée qui y est exprimée pour que soient retrouvés les documents susceptibles d'être pertinents. Cette reformulation peut aussi bien être dans la

même langue (synonymes, hyponymes, etc.) que dans des langues différentes, et pour ce faire, le système du LIC2M utilise des dictionnaires de reformulation monolingue et bilingue.

2.4 Calcul de la proximité sémantique

Le comparateur sert à calculer la proximité sémantique entre la requête et les documents indexés à partir des mots communs (mots de l'intersection requête/documents). Ce comparateur consiste, d'une part, à identifier les meilleures intersections requête/documents, et d'autre part, à regrouper les intersections identiques et leur attribuer un poids. Le résultat est présenté sous forme d'une liste de classes d'intersections dans un ordre croissant de poids. Les documents de la base sont indexés et stockés dans des fichiers inversés. On construit un index pour chacune des langues des documents constituant le corpus et on applique l'analyse linguistique pour les documents à indexer et pour les requêtes.

2.5 Résultats de la recherche cross-lingue

Le système de recherche cross-lingue d'information du LIC2M utilise le modèle booléen pondéré. Lorsque la requête est effectuée, les documents sont renvoyés groupés par classes, qui représentent la répartition des mots de la requête dans les bases de données. Chaque classe contient une liste de documents classés par ordre de pertinence. Par exemple, le moteur de recherche retourne 12 classes pour la requête إدارة موارد المياه "gestion des ressources en eau" (Tableau 1).

Classe	Termes de la requête	Nombre de termes de la requête
1	إدارة_موارد_مياه	1
2	موارد_مياه ,إدارة_موارد	2
3	مياه ,إدارة_وارد	2
4	إدارة ,موارد_مياه	2
5	إدارة_موارد	1
6	موارد_مياه	1
7	إدارة ,موارد ,مياه	3
8	إدارة ,مياه	2
9	إدارة ,موارد	2
10	موارد ,مياه	2
11	مياه	1
12	موارد	1

Tableau 1 : Classes retrouvées pertinentes pour la requête إدارة موارد المياه

3 Alignement de phrases basé sur la recherche cross-lingue

L'alignement de phrases à partir de textes bilingues basé sur la recherche cross-lingue d'information consiste à construire une base de données des phrases du texte cible ($Corpus_{Fr}$) et considérer chaque phrase du texte source ($Corpus_{Ar}$) comme une requête à cette base de données (Figure 1).

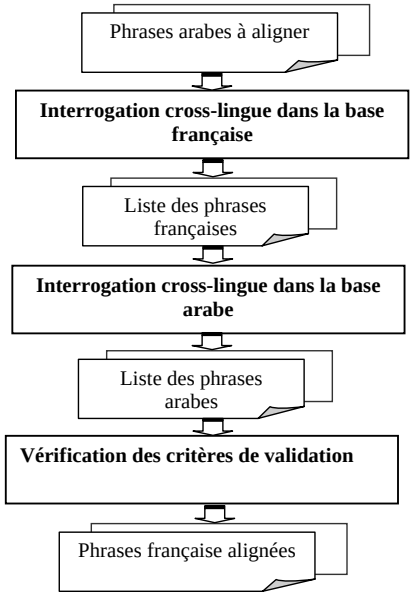


Figure 1 : Etapes de l'alignement de phrases

La validation de l'alignement est basée sur trois critères:

- Position de la phrase dans le document : L'alignement est validé si le rang (numéro d'ordre) de la phrase à aligner se situe dans une fenêtre de tolérance de 10 (rangs) par rapport à la dernière phrase alignée. Cette valeur a été établie expérimentalement.

- Le nombre de termes communs entre la phrase source et la phrase cible (intersection sémantique) doit représenter plus de 50% du nombre des termes de la phrase source.

- Le rapport entre la taille (exprimée en nombre de caractères) de la phrase cible et la taille de la phrase source doit être supérieur à 1.1. La valeur de ce rapport a été fixée expérimentalement. Elle repose sur l'idée qu'*une phrase aura tendance à être traduite par une phrase longue si elle est longue, et par une phrase courte si elle est courte*.

Le processus d'alignement se déroule en quatre étapes :

1. Alignement 1-1 Exact Match: L'objectif est d'obtenir un alignement avec une précision maximale en utilisant les trois critères de validation.
2. Alignement 1-2: Cet alignement consiste à trouver pour la phrase à aligner deux phrases en langue cible en utilisant comme référence le rang de la phrase précédente déjà alignée. Nous utilisons pour valider cet alignement les deux premiers critères.
3. Alignement 2-1: L'objectif de cet alignement est de trouver pour les deux phrases en langue source suivant une phrase déjà alignée une phrase en langue cible en utilisant comme référence le rang de la phrase précédente déjà alignée. Cet alignement est validé par les deux premiers critères.
4. Alignement 1-1 Fuzzy Match: Cette étape consiste à aligner la phrase source avec la première phrase cible de la première classe retournée par le moteur de recherche cross-lingue. Cet alignement n'utilise aucun critère de validation.

Nous décrivons ci-après l'algorithme de l'aligneur 1-1 Exact Match qui constitue la base des autres aligneurs. Cet algorithme utilise les fonctions de l'API du moteur de recherche:

```
- PerformCrosslinguageSearch(Requête, Corpus, Langue source, Langue cible):
retourne l'ensemble des classes retrouvées pertinentes pour la question "Requête"
dans la base de données textuelles "Corpus". Chaque classe est composée d'un
ensemble de phrases dans la langue cible.

- GetNumberOfCommonWords(Classe): retourne le nombre de termes communs entre la
phrase source et la phrase cible (intersection sémantique).

- GetNumberOfWords(Phrase): retourne le nombre de mots pleins d'une phrase.

- GetNumberOfCharacters(Phrase): retourne le nombre de caractères d'une phrase.

Fonction GetExactMatchOneToOneAlignments(Corpus_Ar, Corpus_Fr)
 Pour chaque phrase arabe P_jAr (de rang j) ∈ Corpus_Ar faire
  C_Fr = PerformCrosslinguageSearch(P_jAr, Corpus_Fr, Ar, Fr)
  R = 0; Initialisation du rang de la dernière phrase alignée.
  Pour chaque classe C_lFr ∈ C_Fr faire
   Pour chaque phrase française P_mFr (de rang m) ∈ C_lFr faire
    C_Ar = PerformCrosslinguageSearch(P_mFr, Corpus_Ar, Fr, Ar)
    Pour chaque classe C_qAr ∈ C_Ar faire
     Pour chaque phrase arabe P_qAr ∈ C_qAr faire
      Si P_qAr = P_jAr alors
       NM_Fr = GetNumberOfCommonWords(C_lFr); NM_Ar = GetNumberOfWords(P_jAr);
       NC_Ar = GetNumberOfCharacters(P_jAr); NC_Fr = GetNumberOfCharacters(P_mFr)
       Si (NM_Fr >= NM_Ar/2) et (R – 5 <= m <= R + 5) et (NC_Fr = (1.1)* NC_Ar) alors
        La phrase P_mFr est l'alignement de la phrase P_jAr; R = m
       Fin Si
      Fin Si
     Fin Pour
    Fin Pour
   Fin Pour
  Fin Pour
 Fin Pour
Fin Fonction
```

Par exemple, pour aligner la phrase arabe [4/30] (Phrase de rang 4 dans une base de données contenant 30 phrases) " في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرنية بأن زمن الاحزاب التقليدية قد بلغ نهايته", l'aligneur 1-1 Exact Match procède comme suit:

- La phrase arabe est considérée comme une requête dans la base de données des phrases françaises en utilisant le moteur de recherche cross-lingue. Les phrases retrouvées pertinentes des deux premières classes sont illustrées dans Tableau 2.

- Les réponses de l'interrogation cross-lingue montrent que la phrase française [4/36] est un bon candidat pour l'alignement. Pour confirmer cet alignement, nous utilisons cette phrase comme une requête à la base de données des phrases arabes. Les phrases retrouvées pertinentes pour cette phrase sont groupées dans deux classes dans Tableau 3.

Classe	Nombre de phrases retrouvées	Phrases retrouvées
1	1	[4/36] En Italie, l'ordre des choses a persuadé de manière invisible une majorité d'électeurs que le temps des partis traditionnels était terminé
2	3	[32/36] Au point que, dès avant ces élections, un hebdomadaire britannique, rappelant les accusations portées par la justice italienne contre M. Berlusconi, estimait qu'un tel dirigeant n'était pas digne de gouverner l'Italie, car il constituait un danger pour la démocratie et une menace pour l'Etat de droit [34/36] Après le pitoyable effondrement des partis traditionnels, la société italienne, si cultivée, assiste assez impassible (seul le monde du cinéma est entré en résistance) à l'actuelle dégradation d'un système politique de plus en plus confus, extravagant, ridicule et dangereux [36/36] Toute la question est de savoir dans quelle mesure ce modèle italien si préoccupant risque de s'étendre demain à d'autres pays d'Europe

Tableau 2 : Phrases retrouvées pertinentes pour la phrase arabe à aligner [4/30]

Classe	Nombre de phrases retrouvées	Phrases retrouvées
1	1	[4/30] في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته
2	3	[26/30] يشكل هؤلاء الرجال اكثر ثلاثية مثيرة للسخرية والتقزز في اوروبا، الى درجة ان احدى المجلات الاسبوعية البريطانية اعتبرت في معرض استعادتها للاتهامات القضائية الموجهة الى السيد برلوسكوني قبل هذه الانتخابات ان مسؤولا من هذا النوع "ليس جديرا بحكم ايطاليا" وانه يمثل "خطرا على الديموقراطية" وعلى ""دولة القانون [28/30] وقد تبينت صحة هذه التوقعات المتشائمة، فبعد الانهيار المثير للشفقة للاحزاب التقليدية، شهد المجتمع الايطالي المعروف بثقافته ومن دون ان يبدي حراكا (باستثناء قطاع السينما الذي لجأ الى المقاومة) التدهور الراهن لنظام سياسي يعاني المزيد من الغموض والشطط والسخف والخطورة [30/30] وكل المسألة تكمن في معرفة الى اي مدى يمكن هذا النموذج الايطالي المثير للقلق ان ينتشر غداً في بلدان اوروبية اخرى

Tableau 3 : Phrases retrouvées pertinentes pour la phrase française [4/36]

La première phrase proposée par l'interrogation cross-lingue correspond à la phrase initiale à aligner et plus de 50% des mots sont communs entre les deux phrases. De plus, le rapport

entre la phrase française et la phrase arabe est supérieur à 1.1 et les positions des deux phrases dans les deux bases de données sont identiques. Par conséquent, l'aligneur 1-1 Exact Match considère la phrase française [4/36] comme la traduction de la phrase arabe [4/30].

4 Résultats et Discussions

Pour mener nos expérimentations et être en mesure de calculer la performance de notre aligneur de phrases, nous avons utilisé le corpus MD (Monde Diplomatique) de la campagne ARCADE II (Chiao et al., 2006). Le corpus contient 5 textes arabes (244 phrases) alignés avec 5 textes français (283 phrases).

Pour évaluer l'aligneur au niveau de la phrase, nous avons utilisé les mesures suivantes :

$$\text{Précision} = \frac{|A \cap A_r|}{|A|} \text{ et Rappel} = \frac{|A \cap A_r|}{|A_r|}$$

A correspond à l'ensemble des alignements fournis par l'aligneur et A_r correspond à l'ensemble des alignements corrects.

Les résultats d'alignement sont illustrés dans Tableau 4 et montrent que la précision est autour de 97% et le rappel est autour de 93%. Ces résultats ne prennent pas en compte les alignements partiellement corrects (Alignement 1-1 Fuzzy Match).

Texte parallèle	Précision	Rappel
1	0,969	0,941
2	0,962	0,928
3	0,985	0,957
4	0,983	0,952
5	0,966	0,878

Tableau 4 : Résultats d'alignement au niveau de la phrase du corpus MD

Par ailleurs, l'analyse de ces résultats montre, d'une part, que l'alignement est correct même si le corpus n'est pas parfaitement parallèle, et d'autre part, que la précision dépend fortement des mots discriminants présents dans les phrases source et cible.

5 Conclusion et Perspectives

Nous avons proposé une nouvelle approche pour aligner les phrases d'un corpus parallèle. Cette approche est basée sur une recherche cross-lingue d'information et combine plusieurs sources d'information (dictionnaire bilingue, longueurs des phrases, numéros d'ordre des phrases dans le corpus parallèle). Les résultats que nous avons obtenus montrent des valeurs correctes pour la précision et le rappel même lorsque le corpus n'est que partiellement parallèle. Nos travaux vont maintenant s'étendre, d'une part, à l'utilisation des structures

syntaxiques du corpus parallèle pour améliorer la performance de l'alignement de phrases, et d'autre part, au développement d'un outil d'aide à la traduction basé sur les textes bilingues alignés.

Références

A. ANDREEWSKY, J. P. BINQUET, F. DEBILI, C. FLUHR, B. POUDEROUX. (1981). Le traitement linguistique et statistique des textes et son application dans la documentation juridique. Actes du *Sixième Symposium sur l'Informatique Juridique en Europe*, Thessaloniki, Grèce.

BROWN P., LAI L., MERCIER L. (1991). Aligning Sentences in Parallel Corpora. Actes de *ACL-1991*.

CHIAO Y. C., KRAIF O., LAURENT D., NGUYEN T., SEMMAR N., STUCK F., VÉRONIS J., ZAGHOUANI W. (2006). Evaluation of multilingual text alignment systems: the ARCADE II project. Actes de *LREC-2006*.

DEBILI F. SAMMOUDA E. (1992). Appariement des Phrases des Textes Bilingues. Actes du *14th International Conference on Computational Linguistics*.

DEBILI F., ZOUARI L. (1985). Analyse morphologique de l'arabe écrit voyellé ou non fondée sur la construction automatique d'un dictionnaire arabe. Actes de *Cognitiva-1985*.

FLUHR C., BISSON F., ELKATEB F. (2000). Parallel text alignment using cross-lingual information retrieval techniques. *Parallel text processing*. Kluwer, Boston.

GALE W.A. CHURCH K. W. (1991). A program for aligning sentences in bilingual corpora. Actes du *29th Annual Meeting of Association for Computational Linguistics*.

GAUSSIER E. (1995). *Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues*. Ph.D. Thesis, Paris VII University.

GREFENSTETTE G., SEMMAR N., ELKATEB-GARA F. (2005). Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications. Actes de *ACL-2005*, 31-38.

MELAMED I. D. (1996). A Geometric Approach to Mapping Bitext Correspondence. Actes de *Conference on Empirical Methods in Natural Language Processing*.

SEMMAR N., ELKATEB-GARA F., LAIB M., FLUHR C. (2005). A Cross-language information retrieval system based on linguistic and statistical approaches. Actes du *Deuxième Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la Langue*.

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 orales)

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 1

Utilisation des ontologies pour la modélisation logique d'une commande en langue naturel

Laurent MAZUEL
LIP6, 104 avenue du Président Kennedy, 75016 Paris
laurent.mazuel@lip6.fr

Résumé. Dans cet article, nous nous intéressons à l'interprétation de commandes en langue naturelle pour un agent artificiel. Notre architecture repose sur une modélisation logique de la commande pour l'interprétation sémantique, qui permet de capturer la « structure fonctionnelle » de la phrase, c'est-à-dire les rôles des termes les uns par rapport aux autres. Cet article décrit une méthode d'analyse structurelle de surface qui s'appuie sur l'ontologie de l'agent pour construire cette modélisation logique. Nous définissons tout d'abord un algorithme d'ancrage des termes de la commande dans l'ontologie de l'agent puis nous montrons comment s'en servir pour l'analyse de surface. Enfin, nous expliquons brièvement comment notre modélisation peut être utilisée au moment de l'interprétation sémantique des commandes.

Abstract. In this paper, we focus on natural language interaction for artificial agents. Our architecture relies on a command logical model to enhance the semantic interpretation. It allows us to catch the « functional structure » of the user sentence, *i.e.* each terms compared to each others. This paper describes a partial structural approach which relies on the agent ontology to build a logical form of the sentence. We first define an algorithm to anchor a word from the command in the ontology and we use it to make our partial analysis. Lastly, we explain briefly how to use our model for the semantic interpretation of the user command.

Mots-clés : commande en langue naturelle, analyse structurelle de surface, modélisation logique, ontologies.

Keywords: natural language command, partial structural analysis, logical form, ontologies.

1 Introduction

Dans les applications de commandes en langue naturelle, l'utilisation d'un analyseur syntaxique basé sur des règles grammaticale fortes de la langue pose des problèmes d'efficacité (Milward, 2000; Sabouret & Mazuel, 2005). En effet, les utilisateurs emploient plus régulièrement des mots clés plutôt que des phrases bien structurées (*e.g.* « *drop object low* » ou « *take blue* »). De plus, dans le cadre d'applications réelles, la complexité, la difficulté d'écriture de règles non-spécifiques et de maintenances rendent ces types d'approches complexes à mettre en œuvre et lourdes à utiliser (Sabah, 2006). D'un autre coté, l'utilisation d'un modèle « sac de mots » est insuffisante, générant des problèmes de modélisation impossible à interpréter par la suite (par exemple, « go from London to Boston » et « go from Boston to London » sont représentées par le même sac de mots). C'est pourquoi la majorité des travaux actuels (Hobbs *et al.*, 1997;

Eliasson, 2007) cherchent à effectuer une analyse partielle (ou de surface), afin de réduire le coût de développement, augmenter la portée utilitaire et éviter les écueils des deux modélisations extrêmes précédemment décrites.

Les méthodes actuelles d'analyse de surface s'orientent ainsi vers une modélisation basée sur la logique du premier ou second ordre (Shapiro, 2000; Milward, 2000). Cette modélisation permet à la fois de s'affranchir d'une analyse syntaxique lourde et de conserver suffisamment d'information pour être applicable facilement au moment de l'analyse sémantique. Néanmoins, le défaut de ces systèmes réside dans la définition de ces prédicats, qui doit souvent se faire dans un langage contraint dépendant d'un ensemble d'axiomes logiques spécifiques (Shapiro, 2000; Sadek *et al.*, 1997). Au contraire, l'utilisation d'ontologies dans les systèmes de dialogue a pour objectif de rendre les systèmes plus indépendants de l'application. Elles sont utilisées par exemple pour l'interprétation sémantique d'une commande pour le système (Milward & Beveridge, 2003; Flycht-Eriksson, 2003) et *avant* cette interprétation pour désambiguïser les termes d'une commande (Porzel *et al.*, 2003; Resnik, 1995). Nous pensons qu'il est aussi possible d'exploiter le contenu de l'ontologie pour construire la représentation structurelle logique de la commande, ce qui permet de s'affranchir de la définition de règles dans un langage spécifique.

Dans cet article, nous proposons de définir une méthode d'analyse structurelle de surface pour construire une modélisation logique de la commande basée sur l'étude des concepts et des relations définis dans l'ontologie de l'agent. Notre analyse s'appuie sur un ancrage des termes de la commande dans l'ontologie (nous nous plaçons dans le cadre de l'hypothèse de connectivité sémantique de Sadek (Sadek *et al.*, 1997), qui suppose que tous les concepts de toutes commandes apparaissent dans l'ontologie). En fonction des rôles des termes dans l'ontologie (relation ou classe), nous construisons une représentation de la commande sous forme de prédicats (correspondant aux relations) et d'arguments (instances de classes).

La section suivante présente brièvement notre système d'interprétation de commandes en langue naturel. Nous décrivons l'architecture principale et l'articulation entre les différents composants. La section 3 décrit plus précisément l'ancrage des termes utilisateurs à l'ontologie, l'algorithme de construction logique de la commande et l'interprétation sémantique.

2 Architecture du système de commande en langue naturel

Notre architecture est basée sur le modèle classique des « modules réseaux communicants » (Allen *et al.*, 2000; Seneff, 2002). Cette structure permet le backtrack entre les différents composants ainsi que les réponses anticipées en fonction de l'état du dialogue (figure 1). Nous donnons dans cette section les grandes lignes des modules de l'architecture, en gardant les détails de l'analyse logique et l'interprétation sémantique (comme illustration de l'utilisation de notre analyse) pour la section 3.[1]

[1] L'architecture est définie plus en détails dans (Mazuel & Sabouret, 2006). Elle est utilisée pour la définition d'agents conversationnels sur le web : `http://www-poleia.lip6.fr/~sabouret/demos`.

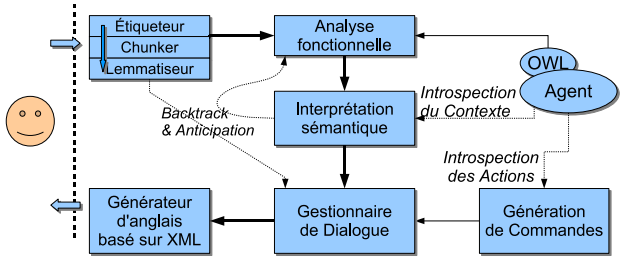


FIG. 1 – Architecture générale

2.1 Analyse morphologique et lexical

Notre module morphologique et lexical est basé sur la bibliothèque d'outils OpenNLP[2]. Nous utilisons les modules *Maximum-Entropy Tokenizer* et *Chunker*, l'étiqueteur et le lemmatiseur basé sur WordNet. L'étiqueteur, le tokenizer et le chunker sont entraînés sur des données anglaises du Wall Street Journal et du corpus Brown. Le dernier modèle proposé est annoncé à 96% d'étiquetage correct sur des données hors base d'apprentissage. Une étude comparative avec le TreeTagger[3] sur quelques exemples tirés de notre application n'a pas montré de pertes très significatives. Le lemmatiseur basé sur WordNet permet la découverte des mots composés de la commande, dans la mesure où le terme existe en tant qu'un des mots d'un *synset* (*e.g.* « *dark red* », « *extra large* »). Nous n'avons pas utilisé le module de résolution d'anaphore de OpenNLP, car elles n'apparaissent que très rarement dans une commande (à la différence de textes longs ou de dialogues).[4]

2.2 Principe de la génération de commandes formelles

Notre système de compréhension des commandes en langue naturelle repose sur une approche ascendante (*i.e.* bottom-up) comme il est possible d'en voir dans (Paraiso & Barthès, 2004). Cette approche utilise une liste *préétablie* de compétences (formelles) et essaye de relier la commande en LN à (au moins) une compétence. Cependant, elles présentent des problèmes d'efficacité en pratique (*e.g.* écriture des compétences, difficulté d'évolution, etc.) qui font que nous utilisons actuellement une version *ascendante générative* basée sur une analyse du code de notre agent (Mazuel & Sabouret, 2006). Notre modèle agent, appelé VDL, permet en effet un accès *à l'exécution* à l'ensemble du code et de son état courant (Sabouret & Sansonnet, 2001). L'algorithme de génération des commandes formelles est inspiré des travaux sur la validation de logiciel par l'analyse des préconditions d'activation d'une action.

Le principe général de l'approche ascendante générative est d'apparier les termes de la commande utilisateur avec les commandes formelles (*i.e.* notées *évènements* en VDL) générées, qui correspondent aux commandes que l'agent est capable de traiter. Cet appariement est le résultat

[2] `http://opennlp.sourceforge.net/`

[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`

[4] En fait, elles apparaissent uniquement lors des dialogues avec l'utilisateur (exemple : « prend le carre vert », « ok », « pose le en haut »). C'est alors le gestionnaire de dialogue qui est responsable de leur résolution (cf. section 2.3).

de l'interprétation sémantique, dont nous parlerons brièvement en section 3.4. A l'issue de cette interprétation sémantique, chaque évènement est associé à un score d'appariement évaluant la proximité de l'événement avec la commande de l'utilisateur. L'objectif de cet article n'est pas de présenter l'algorithme de calcul de ce score (le lecteur intéressé le trouvera dans (Mazuel & Sabouret, 2006)), mais de présenter l'analyse structurelle de surface qui le rend possible.

2.3 Le gestionnaire de dialogue

A l'issue de l'interprétation sémantique le gestionnaire de dialogue utilise le score d'appariement pour déterminer la stratégie de dialogue. Nous utilisons pour cela un système de seuil inspiré de celui proposé par Patty Maes (Maes, 1994) qui permet de faire la différence entre les commandes parfaitement comprises, les commandes incertaines et les commandes non-comprises. Nous avons en plus pris en compte le cas des commandes possibles ou impossibles dans l'état courant de l'agent (Mazuel & Sabouret, 2006).

Pour répondre à l'utilisateur, le gestionnaire de dialogue utilise un générateur d'anglais qui transforme une réponse formalisée en VDL en une phrase anglaise. L'algorithme actuel est très simple et ne produit pas des réponses grammaticalement correctes, mais donne suffisamment d'informations (*i.e.* de mots clefs) pour aider l'utilisateur à reformuler sa commande. Notre objectif à long terme est d'utiliser un générateur performant basé sur XML et les ontologies.

Par exemple, dans le cas d'une ambiguïté, le gestionnaire de dialogue propose à l'utilisateur l'ensemble des commandes possibles dans le contexte courant et utilise le générateur d'anglais pour transformer les commandes formalisées :

```
- I want to go to Boston today.
- Your command is imprecise. I can either :
  - Go Boston with flight is AF1345 and departure time is 8h47
  - Go Boston with flight is AA6543 and departure time is 10h34
```

3 Analyse fonctionnelle logique

Nous décrivons dans cette section comment nous construisons un modèle de la commande de l'utilisateur sous la forme d'un ensemble de prédicats. Nous décrirons d'abord le modèle d'ontologie utilisé, l'algorithme d'ancrage d'un mot dans l'ontologie, puis enfin la construction complète de la modélisation logique de la commande.

Dans la suite de l'article, nous noterons St l'ensemble des chaînes de caractères et pour tout ensemble E, nous noterons $\mathcal{P}(E)$ l'ensemble des sous-ensembles de l'ensemble E.

3.1 Modèle de l'ontologie

Dans notre modèle, l'ontologie d'un agent[5] est un couple $\mathcal{O} = \langle \mathcal{C}, \mathcal{R} \rangle$ dans lequel :

– $\mathcal{C}$ est l'ensemble des concepts (ou *classes*). Un concept représente un ensemble d'objets réuni par les mêmes propriétés. Tout concept $c \in \mathcal{C}$ est caractérisé par un *label* l_c (nous nous

[5]Nous utilisons Jena et OWL pour l'implémentation.

limiterons à un unique label pour simplifier, mais il peut y en avoir plusieurs dans le cas de synonymie, à la manière des synsets de WordNet).
– $\mathcal{R}$ est un ensemble de relations *binaires*. Chaque relation $r \in \mathcal{R}$est caractérisée par un label de relation l_r et un ensemble de couples $E_r \subset \mathcal{C}^2$.

Par soucis de simplification, nous identifierons l_r et l_c respectivement au concept c et à la relation r, et nous noterons ainsi abusivement $\mathcal{C}$ et $\mathcal{R}$ les ensembles de labels de concepts et de relations. Nous noterons $\mathcal{L} = \mathcal{C} \cup \mathcal{R}$. Enfin, nous noterons $\langle c_1, r, c_2 \rangle \in \mathcal{O}$ lorsque les concepts c_1 et c_2 sont reliés par la relation r.

Soulignons que l'ontologie de domaine d'un agent contiendra non seulement les relations usuelles d'hyperonymie (isa) et de meronymie ($partof$), mais aussi des relations plus spécifique du domaine comme $isLargerThan$ ou $leftOf$.

3.2 Ancrage d'un mot dans l'ontologie

Soit W l'ensemble ordonné $w_1, ..., w_n$ des mots utilisés dans la commande. L'ancrage dans l'ontologie consiste à trouver le label l_c ou l_r « le plus proche » pour chaque mot w_i. Notre algorithme se décompose en trois étapes :

1. La simplification morphologique.
2. La recherche des « approximations sémantiques ».
3. L'ancrage proprement dit.

La simplification morphologique consiste à unifier l'écriture des mots ou des groupes de mots (accents, minuscule/majuscule, remplacement des espaces par « _ », *etc*). Par exemple, le terme *bigger* de la commande peut correspondre aux labels *bigger-than*, *is-bigger* encore *biggerThan* selon la notation adoptée dans l'ontologie. Nous ne détaillerons pas le calcul de cette fonction que nous noterons $app_m : St \longrightarrow \mathcal{P}(\mathcal{L})$. Elle prend en entrée un terme de la commande et renvoie la liste de candidats morphologiquement proche parmi les labels présent dans l'ontologie.

La recherche des « approximations sémantiques ».consiste à trouver l'ensemble des termes de l'ontologie les plus proches sémantiquement d'un mot de la commande, en utilisant des mesures de similarité sémantique comme décrites dans (Budanitsky & Hirst, 2006). Nous ne faisons pas ici d'interprétation sémantique de la commande dans le contexte de l'application (nous ne sommes pour l'instant que dans l'analyse structurelle), mais nous cherchons les concepts représentant le mieux les mots utilisés par l'utilisateur. Cette démarche est équivalente aux travaux visant à désambiguïser l'ensemble des concepts reconnus pour un mot d'une commande pour ne choisir que le plus représentatif du contexte de la phrase[6] (Porzel *et al.*, 2003; Resnik, 1995). Par un exemple, dans la commande « buy a place for the Pink Floyd show at the cheapest price », le terme « cheapest » est proche du label de relation *lowerThan* et le le terme « show » du label de concept *concert*[7].

[6]Il n'est pas forcément évident que les phrases employées au sein de notre application correspondent exactement aux sens enregistrés dans WordNet, surtout lorsqu'il s'agit d'un domaine technique (Resnik, 1995). Néanmoins, nous ne nous servons pas de WordNet pour l'interprétation sémantique mais pour *aider* à retrouver les mots de l'utilisateur dans l'ontologie. Ainsi, si le domaine est technique, l'ontologie le sera aussi et la plupart des termes utilisateurs seront retrouvés directement (ou par simplification morphologique). Nous n'avons d'ailleurs pas constaté en pratique de faux-sens à ce niveau.

[7]Ces exemples sous-entendent que « cheapest » et « show » ne sont pas définie dans l'ontologie et n'ont pas d'équivalent morphologique.

Pour cette recherche, nous avons choisi d'utiliser la formule de Jiang & Conrath (Jiang & Conrath, 1997) appliquée aux calculs de probabilités définies par N. Seco (Seco *et al.*, 2004). Cette formule calcule sur WordNet un score de similarité sémantique compris entre $[0, 1]$. Nous ne détaillerons pas cette formule ici car ce n'est pas l'objectif de cet article. Elle a été plusieurs fois évalué et présente les meilleurs résultats actuels en la matière (Budanitsky & Hirst, 2006). Nous noterons $sim_{JC}(w_1, w_2)$ le score de similarité sémantique entre les mots $w_1 \in St$ et $w_2 \in St$.

Nous noterons $app_s : St \longrightarrow \mathcal{P}(\mathcal{L})$ la fonction calculant l'ensemble des labels les plus proches du terme de l'utilisateur. Nous la définissons de la façon suivante :

$app_s(w) =$

$$\begin{cases} \emptyset & \text{si } max^w_{sim} < t_o \\ \{l \in \mathcal{L} \text{ tq } sim_{JC}(l, w) = max^w_{sim}\} & \text{sinon} \end{cases}$$

avec $t_o \in [0, 1]$ le ***seuil d'acceptabilité*** et la similarité maximum $max^w_{sim} = \max_{l \in \mathcal{L}} sim_{JC}(l, w)$. Le seuil d'acceptabilité t_o permet de décider si l'appariement est acceptable ou non[8]. La similarité max^w_{sim} est le score maximal obtenu pour le mot w lors du calcul de similarité sur l'ontologie. Autrement dit, $app_s(w)$ donne l'ensemble des concepts de l'ontologie de similarité maximale avec w.

Ainsi, nous pouvons définir l'ancrage $\mathcal{A} \subset St \times \mathcal{L}$ des mots $w_1, ..., w_n$ dans l'ontologie $\mathcal{O}$:

$\mathcal{A} =$

$$\bigcup_{w \in W} \begin{cases} \bigcup\limits_{l \in app_m(w)} \langle w, l \rangle & \text{Si } app_m(w) \neq \emptyset \\ \bigcup\limits_{l \in app_s(w)} \langle w, l \rangle & \text{Sinon} \end{cases}$$

Soulignons qu'un même terme peut être ancré à plusieurs labels de l'ontologie, donc à plusieurs concepts et/ou relations.

Soulignons aussi que l'interprétation sémantique (cf. section 3.4) utilise les scores calculés à cet étape par sim_{JC} pour déterminer l'imprécision globale de la commande. Cette imprécision est ensuite utilisée par le gestionnaire de dialogue pour déterminer la meilleure stratégie de dialogue.

3.3 Construction des prédicats

Notre objectif est de définir une modélisation logique qui capture la structure fonctionnelle de la phrase, c'est-à-dire de construire un ensemble de prédicats représentant les relations entre les concepts (au sens de l'ontologie $\mathcal{O}$) tels qu'ils sont exprimées dans la commande. Par exemple, dans « *the big object next to the book* », l'utilisateur exprime une relation « *next-to* » entre « *big object* » et « *book* ».

Pour cela, chaque terme est considéré du point de vue de son ancrage dans l'ontologie : si c'est une relation, nous la modéliserons sous la forme d'un prédicat et nous devons rechercher ses arguments dans la commande parmi les autres termes/concepts. En adoptant une représentation

[8] Actuellement et empiriquement, la valeur du seuil d'acceptabilité t_o est de 0.7.

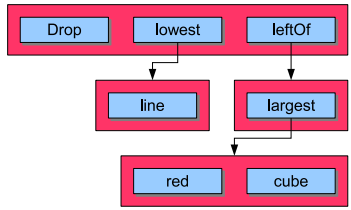


FIG. 2 – Modélisation de « drop on the lowest line, left of the largest red cube »

arborescente des prédicats, les nœuds des arbres sont les termes de la commande. Les termes qui sont des labels de concepts sont représentés par des feuilles. Les termes qui sont des labels de relations sont représentés par des nœuds dont les fils sont les arguments de la relation dans la commande de l'utilisateur. Par exemple, dans la phrase « *drop on the lowest line, left of the largest cube* », « *drop* », « *line* », « *red* » et « *cube* » sont des feuilles, « *lowest* » aura comme fils « *line* ». Nous obtenons alors le résultat présenté dans sur figure 2.

Toute la difficulté de cette construction réside dans la capacité à déterminer quel terme est un argument de quelle relation. Idéalement, nous devrions nous appuyer sur l'analyse sémantique de la phrase et sur les définitions des relations dans l'ontologie pour identifier les instances correspondant à des arguments de l'agent, en utilisant du *backtrack* pour rechercher toutes les permutations possibles.

Mais dans un premier temps, par soucis d'efficacité, nous utiliserons l'heuristique suivante, tirée de nos observations sur les relations dans la langue anglaise :

> Les arguments d'une relation sont soit l'ensemble des termes restant dans le syntagme nominal de la relation, soit dans l'ensemble des termes du syntagme immédiatement suivant.

La force de cette heuristique est qu'elle prend aussi en compte le traitement des comparatifs et des superlatifs :

1. Si un superlatif apparaît, il l'est alors à titre d'adjectif descriptif de l'objet. Les termes de la commande reliés appartiennent donc au *même syntagme* (*e.g.* « *the biggest square* », « *the darkest big object* », etc.).
2. Si un comparatif apparaît, l'objet de la comparaison est séparé par l'utilisation d'une conjonction (« *than* », etc.) et donc dans le syntagme suivant (*e.g.* « *higher than the cube* », « *left to the current position* », etc.).

Formellement, soit S l'ensemble *ordonné* $\{c_1, c_2, ..., c_n\}$ composé de n *chunks* tel que $\forall i \in [1, n]$, $c_i = \{s_{i,1}, s_{i,2}, ..., s_{i,k_i}\}$ où les $s_{i,j}$ sont les termes de la commande utilisateur, regroupés en *chunks*[9]. La fonction $\tau : S \longmapsto S_a$ construit l'ensemble d'arbres S_a à partir de la modélisation de la commande chunkée S. Les éléments de S_a seront représentés en utilisant une notation prédicat/valeurs (chaque prédicat représentant un nœud, et ses valeurs les fils du nœud).

La fonction τ est définie récursivement par : $\tau(S) =$

[9]La commande de l'utilisateur est l'ensemble ordonné $S_{user} = \{s_{1,1}, s_{1,2}...., s_{1,k_1}, s_{2,1}, ..., s_{2,k_2}, ..., s_{n,1}, ..., s_{n,k_n}\}$.

$$\begin{cases} \{s_{1,1}\,(\tau(\{\{s_{1,2},...s_{1,k_1}\}))\} \cup \tau(\{c_2,...,c_n\}) & \text{si } (k_1 > 1) \wedge (\exists c \in \mathcal{R}.\, \langle s_{i,j}, c\rangle \in \mathcal{A}) \\ \{s_{1,1}\,(\tau(\{c_2\}))\} \cup \tau(\{c_3,...,c_n\}) & \text{si } (k_1 = 1) \wedge (\exists c \in \mathcal{R}.\, \langle s_{i,j}, c\rangle \in \mathcal{A}) \\ \{s_{1,1}\} \cup \tau(\{\{s_{1,2},...s_{1,k_1}\}, c_2,...,c_n\}) & \text{sinon} \end{cases}$$

avec $\tau(\emptyset) = \tau(\{\emptyset\}) = \emptyset$. Autrement dit, l'arbre S_a est obtenu en transformant chaque relation de la commande en nœud dont les fils sont les termes restant du chunk (lorsque $k_1 > 1$) ou les éléments du chunk immédiatement suivant lorsque la relation est le dernier élément du chunk ($k_1 = 1$). Les concepts sont systématiquement transformés en feuilles. Pour mieux comprendre cette opération, considérons l'exemple suivant : « *drop on the lowest line, left of the largest red cube* » est chunkée en :

> [VP Drop :VB] [PP on :IN] [NP the :DT lowest :JJS line :NN] [? ? , :,] [NP left :NN] [PP of :IN] [NP the :DT largest :JJS red :JJ cube :NN]

Après filtrage des termes non- significatifs, nous obtenons l'ensemble d'ensembles :

$$S = \{\{drop\}, \{lowest, line\}, \{leftof\}, \{largest, red, cube\}\}$$

Nous obtenons alors $\tau(S) = \{drop, lowest(line), leftof(largest(red, cube))\}$, représenté sous forme d'arbre sur la figure 2.

3.4 Interprétation sémantique

L'analyse fonctionnelle décrite précédemment (*cf.* figure 3) permet :

1. La construction d'un ensemble d'arbres représentant la commande ;
2. L'ancrage de cet arbre, par l'ancrage de chacun de ses termes, sur l'ontologie.

Ces deux propriétés sont à la base de notre modèle d'analyse sémantique. En effet, de manière similaire, nous ancrons semi-automatiquement le code de l'agent VDL sur l'ontologie au moment de l'écriture de l'agent. Ainsi, les évènements formels construits par notre algorithme ascendant génératif, utilisant des termes issus du code VDL, sont déjà ancrées dans l'ontologie (chaque commande générée ayant un ancrage diffèrent). Nous nous retrouvons alors dans une situation proche d'un problème d'appariement d'ontologies selon une ontologie de référence (*e.g. (Aleksovski* et al.*, 2006)*). L'objectif est alors d'évaluer comparativement ses deux ancrages, afin de pouvoir décider quelles sont les commandes générées les plus proches de la commande en langue naturelle de l'utilisateur.

C'est l'ancrage des termes de la commande dans l'ontologie qui permet de se ramener à un problème (non trivial) d'alignement d'ontologies. En effet, il nous est alors possible de calculer l'alignement demandant le moins « d'effort » d'approximation entre les deux ensembles de termes ancrés et donc d'en déduire quel couple (évènement/structure de commande) est le meilleur candidat comme résultat à cette interprétation sémantique.

La modélisation logique structurée de la commande utilisateur est ensuite utilisée au moment de l'interprétation sémantique pour calculer la fermeture transitive de la relation dans le contexte courant de l'agent. Par exemple, si l'utilisateur parle d'un objet « *à côté du livre* », notre interprétation sémantique donne l'ensemble des positions correspondant à « *à côté du livre* » en fonction de la position du livre dans l'état courant.

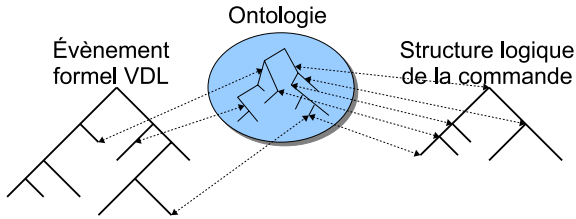


FIG. 3 – État après la modélisation logique avec un évènement formel et un seul arbre de la commande.

4 Conclusion

Dans cet article, nous proposons un algorithme de modélisation d'une commande sous la forme d'un ensemble de propositions logiques qui s'appuie sur l'utilisation de l'ontologie de l'agent. Les symboles de prédicats utilisés sont directement extrait à partir des termes de la commande en fonction de leur proximité avec les concepts de l'ontologie. Les rôles de prédicats ou arguments pour chaque terme sont choisis à partir de leur définition dans l'ontologie. Ce mécanisme ne nécessite donc pas l'utilisation d'un formalisme particulier pour définir les règles d'analyse syntaxique. La modélisation obtenue est simple à interpréter et à utiliser, en particulier pour l'interprétation sémantique de la commande. La plupart des systèmes de dialogues actuelles étant basés sur l'utilisation d'ontologies pour l'interprétation sémantique, l'approche est applicable à large échelle sur des systèmes d'implémentation diverses.

La méthode d'ancrage des termes de la commande dans l'ontologie (c'est-à-dire la recherche du concept de l'ontologie le plus proche sémantiquement d'un terme donné) que nous avons présenté repose sur un algorithme de similarité sémantique basé sur WordNet. L'évaluation préliminaire de notre système actuellement en court présente des résultats encourageants. Cependant, nous voudrions la valider sur d'autres agents et ontologies que celles que nous avons utilisées jusqu'à présent, afin de montrer la généricité de notre approche.

Références

ALEKSOVSKI Z., TEN KATE W. & VAN HARMELEN F. (2006). Exploiting the structure of background knowledge used in ontology matching. In *Proc. Workshop on Ontology Matching in ISWC2006* : CEUR Workshop Proceedings.

ALLEN J., BYRON D., DZIKOVSKA M., FERGUSON G., GALESCU L. & STENT A. (2000). An architecture for a generic dialogue shell. *NLENG : Natural Language Engineering*, **6**.

BUDANITSKY A. & HIRST G. (2006). Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, **32**(1), 13–47.

ELIASSON K. (2007). Case-Based Techniques Used for Dialogue Understanding and Planning in a Human-Robot Dialogue System. In *Proc. of IJCAI07*, p. 1600–1605.

FLYCHT-ERIKSSON A. (2003). Design of Ontologies for Dialogue Interaction and Information Extraction. In *Proc. Workshop on Knowledge and reasoning in practical dialogue systems (IJCAI'03)*.

HOBBS J., APPELT D., BEAR J., ISRAEL D., KAMEYAMA M., STICKEL M. & TYSON M. (1997). FASTUS : A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. *Finite-State Language Processing*, p. 383–406.

JIANG J. & CONRATH D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. on International Conference on Research in Computational Linguistics*, p. 19–33, Taiwan.

MAES P. (1994). Agents that reduce workload and information overload. *Communications of the ACM*, **37**(7), 30–40.

MAZUEL L. & SABOURET N. (2006). Generic command interpretation algorithms for conversational agents. In *Proc. Intelligent Agent Technology (IAT'06)*, p. 146–153 : IEEE Computer Society.

MILWARD D. (2000). Distributing representation for robust interpretation of dialogue utterances. In *ACL*, p. 133–141.

MILWARD D. & BEVERIDGE M. (2003). Ontology-based dialogue systems. In *Proc. 3rd Workshop on Knowledge and reasoning in practical dialogue systems (IJCAI03)*, p. 9–18.

PARAISO E. & BARTHÈS J. (2004). Architecture d'une interface conversationnelle pour les agents assistants personnels. In P. PAROUBECK & J.-P. SANSONNET, Eds., *Actes de la Journée d'Etude ATALA Agental « Agents et Langue »*, p. 83–90, Paris, France : ATALA ATALA.

PORZEL R., GUREVYCH I. & MULLER C. (2003). Ontology-based contextual coherence scoring. In *Proc. of the Fourth SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan.

RESNIK P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, p. 448–453.

SABAH G. (2006). *Compréhension des langues et interaction*. Cognition et Traitement de l'Information. Hermes-Lavoisier.

SABOURET N. & MAZUEL L. (2005). Commande en langage naturel d'agents VDL. In *Proc. 1st Workshop sur les Agents Conversationnels Animés (WACA)*, p. 53–62.

SABOURET N. & SANSONNET J. (2001). Automated Answers to Questions about a Running Process. In *Proc. CommonSense 2001*, p. 217–227.

SADEK D., BRETIER P. & PANAGET E. (1997). Artimis : Natural dialogue meets rational agency. In *IJCAI (2)*, p. 1030–1035.

SECO N., VEALE T. & HAYES J. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proc. ECAI'2004, the 16th European Conference on Artificial Intelligence*, p. 1089–1090.

SENEFF S. (2002). Response planning and generation in the MERCURY flight reservation system. In *Computer Speech and Language*, volume 16, p. 283–312.

SHAPIRO S. (2000). Sneps : a logic for natural language understanding and commonsense reasoning. *Natural language processing and knowledge representation : language for knowledge and knowledge for language*, p. 175–195.

L'analyse morphologique des réponses d'apprenants

Alexia Blanchard
LIDILEM – Université Stendhal, Grenoble
Alexia.Blanchard@u-grenoble3.fr

Résumé. Nous présentons une approche empirique de l'évaluation automatique des réponses d'apprenants au sein d'un système d'Apprentissage des Langues Assisté par Ordinateur (ALAO). Nous proposons la mise en place d'un module d'analyse d'erreurs attestées sur corpus qui s'appuie sur des techniques robustes de Traitement Automatique des Langues (TAL). Cet article montre la réalisation d'un module d'analyse de morphologie flexionnelle, en situation hors-contexte, à partir d'un modèle linguistique existant.

Abstract. We present an empirical approach to the automated evaluation of learner's answers in a CALL system (Computer Assisted Language Learning). We suggest the realization of an error parsing module using NLP techniques (Natural Language Processing). The errors stem from a language learners corpus. This article describes the implementation, from an existing linguistic model, of an inflectional context-free morphology parser.

Mots-clés : TAL, ALAO, détection d'erreurs, morphologie flexionnelle, rétroaction(s).

Keywords: NLP, CALL, errors detection, inflectional morphology, feedback.

1 TAL et analyse des réponses en ALAO

L'Apprentissage des Langues Assisté par Ordinateur (ALAO) a pour but le développement d'outils et de ressources destinés à l'apprentissage des langues. Parmi ces outils, nous trouvons les logiciels éducatifs qui proposent un aspect interactif (l'apprenant interagit avec le système), et également, en théorie, un apprentissage individualisé, personnalisé et autonome. Deux apprenants peuvent acquérir les mêmes connaissances à leur rythme, avec un parcours adapté au sein du logiciel. Il existe une typologie de ces logiciels (Lancien, 1997 ; Wyatt, 1988) : les logiciels d'exploration (l'apprenant simule une situation dans le but de rechercher des informations sur le réseau), les logiciels de référence (l'apprenant dispose d'outils pour l'aider dans sa recherche d'informations comme des encyclopédies) et les logiciels de structure (l'apprenant est soumis à une autoévaluation). Nous nous intéressons ici à ce dernier type de logiciels, qui proposent des exercices, généralement structuraux, et plus particulièrement à l'évaluation automatique de ces exercices, qui constitue un des critères de qualité du « potentiel d'utilisation » d'un logiciel d'apprentissage selon Nielsen (1993, p.20) : « *Errors messages should be expressed in plain language (no codes), precisely indicate the*

problem, and constructively suggest a solution[1] ».
Cette évaluation implique la détection des erreurs commises par l'apprenant et leur analyse afin de produire des rétroactions adaptées (feed-back). Certains se sont demandés s'il était possible d'utiliser des correcteurs orthographiques et grammaticaux existants pour corriger les erreurs de l'apprenant. Une étude (Cordier-Gauthier, Dion, 2002) a démontré que ces correcteurs ont tendance à surdétecter par rapport à la correction d'enseignants. De plus, l'analyse des erreurs propose plusieurs solutions pour une même forme, ce qui limite leur utilisation dans un dispositif d'ALAO.

L'incapacité des systèmes d'ALAO à prendre en compte les propriétés intrinsèques à une langue restreint leurs possibilités. Les principales critiques de ces systèmes restent leur fermeture (contenus des exercices prédéfinis et figés), une détection d'erreurs fondée sur de simples comparaisons de chaînes (pattern-matching), et par conséquent des rétroactions / feed-back pédagogiquement peu adaptés à l'apprenant, du type « vrai/faux ». Ces aspects font défaut depuis toujours dans les travaux en ALAO (Antoniadis et al., 2005b).
Nous pensons que l'intégration d'outils de Traitement Automatique des Langues (TAL) peut permettre de débloquer cette situation et d'améliorer l'analyse des réponses (Antoniadis et al., 2006). Le TAL (Traitement Automatique des Langues) est au carrefour entre linguistique et informatique, et ne considère plus la langue comme une simple suite de caractères, mais comme un système à deux niveaux (sens et forme) (Fuchs et al., 1993).

L'utilisation du TAL dans la conception de systèmes d'ALAO n'est certes pas une idée nouvelle. Certains projets comme Freetext, ont développé un logiciel d'ALAO pour des apprenants du français (Granger et al., 2001). Pour ce faire, ils se sont fondés sur l'analyse du corpus FRIDA (French Interlanguage DAtabase), qui a abouti à un étiquetage et une typologie des erreurs. Cette dernière a ensuite permis la création d'un outil de détection d'erreurs orthographiques, syntaxiques et sémantiques. Une évaluation effectuée en 2003 sur 120 phrases a permis de comparer le prototype de FreeText avec le correcteur de Microsoft Word® 2000 (L'Haire, Vandeventer, 2003). Les résultats montrent que le système FreeText est relativement plus performant que Word®. Toutefois, les auteurs s'accordent pour mettre en évidence la surdétection d'erreurs de la part du système Freetext, phénomène qui peut complètement perturber l'apprenant en situation d'apprentissage.
C'est pourquoi nous préférons privilégier une méthode empirique (1) en utilisant des outils fiables du TAL (2) en réalisant un module d'analyse d'un certain type d'erreurs dans le but, à long terme, de réaliser un système plus complet, intégrable dans un système d'ALAO (3) en traitant des erreurs issues d'un corpus.
Un premier travail a donc consisté à élaborer un module d'analyse qui prend en entrée une forme simple (nom, adjectif ou verbe) et qui analyse ses traits morphologiques afin de détecter une éventuelle erreur de morphologie flexionnelle. La sortie propose un ou plusieurs diagnostic(s) pour une même forme, erronée ou non. Par exemple, la forme *plaçe* est analysée comme une forme erronée du verbe placer à la 1ère personne du singulier de l'indicatif, du subjonctif et à la 3ème personne du singulier de l'indicatif, et du subjonctif (4 diagnostics).

Nous verrons dans un premier temps la méthodologie adoptée pour définir les propriétés de cet analyseur, puis nous décrirons les travaux menés au CRISS pour l'analyse et la génération

[1] « Les messages d'erreurs doivent être écrits en langage naturel (pas de codes), ils doivent indiquer précisément le problème, et suggérer une solution qui aidera l'apprenant à rectifier son erreur ».

morphologique dans un cadre autre que l'ALAO ; travaux que nous avons repris et adaptés à notre sujet. Enfin, nous montrerons les premiers résultats du prototype réalisé.

2 Positionnement

Le domaine de l'ALAO est à la croisée de plusieurs disciplines : didactique des langues, linguistique et TAL. La conception d'un système d'ALAO se doit, à notre avis, de prendre en compte les aspects de ces différentes disciplines.
En ce qui concerne l'évaluation automatique des productions écrites de l'apprenant, il est reconnu qu'un système d'ALAO ne peut être didactiquement valide que s'il est capable d'analyser correctement ces productions afin de ne pas générer des rétroactions erronées (Rézeau, 2004). Une maîtrise incomplète de l'analyse pourrait impliquer un apprentissage partiellement voire complètement biaisé.
Pour ce faire, nous avons préféré brider dans un premier temps notre analyseur, en considérant les contraintes des trois domaines concernés. Notre but n'est pas de proposer une analyse exhaustive des réponses d'apprenants en production libre. Il faut être conscient des fortes contraintes didactiques d'un système d'ALAO. En effet, une telle analyse doit permettre de générer automatiquement une, et une seule explication de chaque erreur pour que l'apprenant puisse comprendre et corriger ses lacunes.
Se posent alors plusieurs problèmes : Quelles informations extraites de l'analyse permettent réellement de prédire une explication quant à l'erreur ? Devons-nous traiter toutes les erreurs théoriquement possibles ou devons-nous cibler les erreurs les plus fréquentes chez un apprenant du français ?
Pour répondre à ces questions, nous avons choisi de nous appuyer sur l'outil TAL le plus fiable aujourd'hui, en l'occurrence l'analyse morphologique flexionnelle :
« La question des flexions grammaticales est, depuis longtemps clairement maîtrisée […] », (Fuchs et al., 1993, p. 102).
Toutefois, l'analyse morphologique comme nous la concevons en TAL repose sur un texte correctement orthographié. Une telle analyse est donc différente d'une analyse morphologique des réponses d'apprenants qui a pour but la génération d'une rétroaction. Pour réaliser un tel analyseur, il nous a donc fallu effectuer un certain nombre de choix :

- **Les types d'erreurs** : Si nous nous penchons sur les études antérieures en didactique des langues, des erreurs fréquentes de morphologie flexionnelle se dégagent :

    - Erreurs de surgénéralisation (Fayol, 2001) : cette notion dénote le fait que les apprenants ont tendance à appliquer la règle d'accord du pluriel sur une autre catégorie (exemple : *les attitudent*).

    - Erreurs d'omission de flexions du nombre (Astolfi, 1997).

    - Erreurs de généralisation (Ågren, 2005) : la généralisation concerne les erreurs d'utilisation d'une flexion d'un type de nom par exemple, sur un autre type de nom (exemple : *les lieus*).

Notre modélisation s'appuie sur ces tendances, mais également sur des erreurs extraites du corpus FRIDA[2] mis à notre disposition dans le cadre du projet Idill[3].

- **Le / les diagnostic(s)** : Nous différencions les notions de diagnostic et de rétroactions. Les rétroactions sont les explications proposées à l'apprenant pour lui expliquer son erreur, alors que les diagnostics sont l'ensemble des informations qui résultent de l'analyse morphologique, sachant que plusieurs diagnostics peuvent être associées à une même forme. Nous préférons rester prudents en conservant toutes les interprétations possibles pour une même forme, afin de ne pas risquer de biaiser la qualité de la rétroaction générée à partir de ce / ces diagnostic(s). Chaque forme sera ainsi associée à son lemme, sa catégorie syntaxique, ses traits morphologiques, le type d'erreur morphologique (notre analyseur est capable d'analyser les formes correctes comme les formes erronées). Si nécessaire, la désambiguïsation pourra être faite à un niveau didactique (via le modèle de l'apprenant, le domaine d'apprentissage, etc.).

- **L'algorithme d'analyse** : A partir des deux premiers points énoncés ci-dessus (ce que prend en entrée l'analyseur et ce qu'il génère en sortie), nous pouvons schématiser le fonctionnement du module réalisé de la façon suivante :

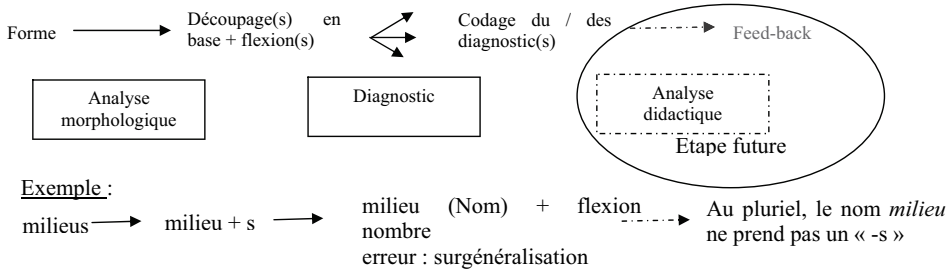


Schéma 1 : Algorithme d'analyse du module et exemple

3 Analyse morphologique des réponses d'apprenants

3.1 L'analyse morphologique en TAL

En TAL, l'analyse morphologique s'appuie sur les règles de formation des mots révélés par les travaux en linguistique. Elle consiste, à partir d'un texte source à extraire des informations pour chaque mot.
Nous trouvons trois écoles principales d'analyse morphologique (Fuchs et al., 1993) qui proposent différentes solutions pour un même problème : comment découper une forme fléchie en base et flexion(s) dans le but de générer un maximum d'informations:

[2] Tous les exemples de formes erronées cités dans l'article sont issus de ce corpus.

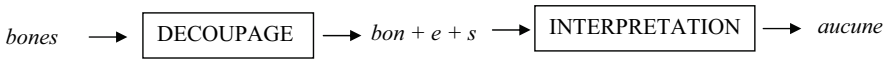
[3] Integrated Digital Language Learning (http ://www.idill.org). Ce projet a été développé dans le cadre du réseau d'excellence européen Kaleidoscope (http://www.noe-kaleidoscope.org).

- **Les dictionnaires de formes fléchies** : Cette façon de procéder a l'avantage de ne demander que peu de traitements informatiques. Elle consiste à lister toutes les formes fléchies des mots du français dans un dictionnaire. Il n'y a donc pas de découpage de forme.

- **Les automates à états finis** : un ensemble d'automates définit les combinaisons autorisées entre bases et affixes contenues dans un dictionnaire. Une forme est validée si elle est acceptée par un automate (e.g. *veuve-> veu +ve*).

- **Le modèle à deux niveaux** : La grammaire du modèle à deux niveaux permet le découpage d'une forme en prenant en compte ces variations phonologiques. L'analyse du mot *veuve* est donc possible, et ce mot est reconnu comme une forme fléchie de la base *veuf*, à laquelle sont rattachées des informations morphologiques (e.g. *veuve->veuf + e*).

Pour notre problématique, la méthode utilisée doit obligatoirement permettre une interprétation fine et pertinente des informations morphologiques dans le but de proposer des rétroactions de qualité. L'utilisation de dictionnaire ne semble pas appropriée car elle ne permet pas de détecter une erreur. Par exemple, *lieus* ne peut être analysé car c'est une entrée inconnue. L'utilisation des automates à états finis limite l'interprétation des erreurs. Si nous prenons l'exemple de *carnavaux*, cette technique découpera cette forme en *carnav + aux*. Ce découpage ne permet pas d'extraire d'informations pertinentes pour notre sujet.
L'analyse à deux niveaux semble donc la plus pertinente pour notre problématique. Toutefois, cette méthode n'est valable que pour les formes correctement orthographiées. Une réadaptation de celle-ci est donc nécessaire pour l'analyse d'erreurs.
Le modèle développé au CRISS (Lallich-Boidin et al., 1990) suit le principe du modèle à deux niveaux et nous semble donc adapté et adaptable pour l'analyse d'erreurs morphologiques car il se fonde sur un découpage des formes qui permet une interprétation linguistique des bases et flexions extraites. Voici un exemple d'analyse avec la forme *bones* :



Il peut découper la forme *bones* mais n'est certes pas capable d'interpréter cette analyse du simple fait qu'il n'a pas été conçu pour reconnaître des formes erronées. Toutefois, le découpage proposé nous permet de modéliser une interprétation de forme erronée.

3.2 L'analyseur morphologique du CRISS

3.2.1 Description générale

Dans les années 90, une équipe du CRISS a développé un analyseur du français en collaboration avec A. Berrendonner (1990). Il prend en entrée un texte écrit, auquel il applique une analyse morphologique et syntaxique, pour en ressortir une représentation logique. Le système a une architecture modulaire, c'est-à-dire qu'analyse morphologique et analyse syntaxique sont séparées. Ceci nous permet d'utiliser facilement les données et algorithmes relatifs à l'analyse morphologique.
Le modèle du CRISS réalise un découpage interprétable de forme (en base et flexion(s)) dans le sens où il utilise les propriétés linguistiques intrinsèques à la langue pour ce qui concerne la formation flexionnelle des mots. Ces propriétés concernent entre autres la place d'apparition des flexions (e.g. base + flexion genre + flexion nombre)).
Ce modèle original utilise un dictionnaire de bases, qui pour chaque base recense le lemme

correspondant, la catégorie syntaxique de ce lemme, le nom du modèle auquel il est associé (dans le cas d'un lemme variable). Voici un exemple d'entrées lexicales :

frère (base), **frère** (lemme)**, F** (Catégorie[4])**, coup** (Modèle de flexion)

Les modèles de flexions permettent de décrire les phénomènes de variations flexionnelles pour un lemme donné. Si nous reprenons l'exemple du lemme *frère*, voici la description du modèle associé :

coup [masculin, singulier] [Ø] [s] masculin, pluriel

Le nom du modèle associé à *frère* est *coup*. La première paire de crochets signifie que la base associée au modèle *coup* est au masculin singulier, la seconde paire vide montre que cette base n'accepte pas de flexion du féminin et qu'aucune flexion n'est nécessaire pour former le masculin à partir de la base, et la troisième paire de crochets signifie que la base *coup* fait son pluriel en *-s*.
Dans le souci de privilégier les calculs aux données[5], l'équipe du CRISS a décidé de ramener au maximum les exceptions aux cas généraux. Cette opération s'effectue à l'aide de règles dites de *régularisations*. Ces dernières ont pour but de réduire le nombre de bases pour un même mot, et donc le nombre d'entrées du dictionnaire. Elles permettent également de diminuer le nombre de flexions et de modèles.
Il existe deux types de régularisations :

- les régularisations portant sur la forme : elles consistent à remplacer le suffixe d'une forme par un autre suffixe, avant le découpage de celle-ci.

- les régularisations portant sur la base : elles interviennent après le découpage d'une forme en base(s) et flexion(s). Le suffixe de la base est substitué par un autre.

Par exemple, pour le mot *travaux*, l'analyseur cherche dans un premier temps la base *travaux* dans le dictionnaire*,* qu'il ne va évidemment pas trouver. Il ramène alors cette forme au cas régulier en substituant *-ails* à la flexion *-aux* (i.e. *travaux* en *travails*). Une régularisation de forme a été effectuée.
A cela, il associe la base *travail* au modèle de flexion adéquate et peut générer les informations suivantes, après découpage :

travail (nom) + s → masculin, pluriel

Pour l'analyse de *plaça*, l'analyseur ne trouvant pas de base *plaç* après découpage en *plaç* et *-a*, il ajoute un *-e* à cette dernière (1ère régularisation de base). La base *plaçe* n'étant toujours pas dans le dictionnaire, il substitue le *çe* en *ce* (2ème régularisation de base) et obtient la base *place*. Les informations liées à cette base permettent de générer les informations suivantes :

place (radical du verbe placer) + a → 3ème personne du singulier du passé-simple.

---

[4] La catégorie F regroupe les noms et adjectifs, la catégorie V les verbes.

[5] Comme le souligne l'équipe du CRISS (Lallich-Boidin et al., 1990), les exceptions nécessitent souvent des données de grande taille par rapport aux cas dits « généraux ». C'est pourquoi ces phénomènes sont transformés des les premiers traitements, afin d'éviter une surcharge de la taille des données de l'analyseur.

3.2.2 Les limites du système pour l'ALAO

Si nous tentons d'analyser des formes morphologiquement erronées avec le modèle tel quel, nous soulevons deux types de problèmes illustrés à travers les exemples suivants :

nationals → national + s → national (adj.) + flexion → *national*, masculin, pluriel, forme correcte de nombre

journeaux → régularisation : journeals → journeal + s → aucune analyse (la base *journeal* n'existe pas dans le dictionnaire)

Ces deux exemples montrent que l'analyseur du CRISS n'a pas été conçu pour analyser des formes erronées. Toutefois, le premier exemple montre que le découpage effectué est pertinent et facilement adaptable à notre problème. Le second exemple met en avant les cas où aucun découpage n'aboutit à une analyse. Nous avons donc modifié le modèle existant afin de pouvoir découper toutes les formes erronées et ainsi les analyser.

3.3 Le module d'analyse morphologique des réponses d'apprenants

Afin d'adapter le modèle du CRISS, il est important de savoir comment il se comporte face à des erreurs, et s'il est capable de les découper en base et flexions. Si ce découpage est possible, alors l'erreur peut être détectée et analysée. Notre préoccupation est donc de modifier le modèle au minimum afin de permettre ce découpage.
Dans un premier temps, nous avons expérimenté la généralisation à travers les données mêmes du modèle, afin de mettre en évidence tous les cas de généralisation possibles au sein du modèle et de mesurer ses potentialités. Les résultats montrent que la totalité des formes obtenues sont découpables.
Nous avons ensuite réalisé un test du système sur les erreurs du corpus FRIDA (57 pour les noms et adjectifs, 90 pour les verbes). Nous les avons tout d'abord classées selon qu'elles sont découpables par les données d'origine du modèle ou non. Ce classement a révélé un taux important pour les erreurs de généralisation et surgénéralisation, autant pour les verbes que pour les noms et adjectifs. Nous avons élargi les types d'erreurs à traiter grâce au corpus qui a mis en avant des cas de sous-généralisation où l'erreur est due à un oubli de régularisation (e.g. *nationals*).

Le principal défaut du modèle du CRISS repose sur le fait qu'il n'associe pas les régularisations avec les bases et flexions adéquates, ce qui entraîne une surgénération d'erreurs analysées comme forme correcte. Par exemple, la forme *carnavaux* sera analysée comme la forme régularisée au pluriel en *–aux* de la base *carnaval*.
Par conséquent, l'une des principales modifications de ce modèle va consister à intégrer les régularisations de forme et de base à l'intérieur même des modèles de flexions.
Ceci implique la modification des modèles existants, mais également la création de nouveaux modèles qui permettent d'intégrer chaque régularisation.
Afin de traiter les cas d'utilisation de flexion et de régularisations erronées[6] (e.g. *journeaux*),

[6] Nous entendons par *erronée* le fait que l'emploi de ce type de flexions et de régularisation fasse échouer l'analyse de ces fautes par l'analyseur non-modifié, car ni ces flexions, ni ces régularisations n'existent dans la langue française.

nous avons ajouté ces régularisations et flexions erronées au modèle. (e.g. la régularisation erronée *–eaux* en *–als* en plus de la régularisation existante *–aux* en *–als*).

Après la modification du modèle, nous obtenons 88 modèles de flexions, contre 61 à l'origine, et 83 règles de régularisations contre 63. Quant au dictionnaire, nous avons modifié de façon semi-automatique les 70 000 entrées afin d'associer les modèles de flexions à chaque base.

Nous trouvions judicieux, en plus des informations « classiques » générées par l'analyse morphologique, de créer une typologie des erreurs de morphologie flexionnelle, en nous appuyant sur le corpus FRIDA et sur le modèle (i.e. mauvaise manipulation des règles de régularisations). Cette typologie permet de proposer une première interprétation, éventuellement utilisable pour la future génération du feed-back. Par exemple, *nationals* est typé comme un oubli d'une régularisation de forme (i.e. *-aux* en *-als*) via le code F1_REGF_O généré lors de son analyse. Quant à *journeaux*, son analyse produit le code F2_REGF, qui correspond à une utilisation d'une régularisation de forme inexistante en français (i.e. *–eaux* en *–als*).

Le module a été réalisé en Prolog II+[7], pour sa capacité à séparer données et algorithmes (grammaire déclarative) et pour son principe de *bactracking* (qui permet une analyse combinatoire et multiple).Voici la sortie proposée par notre analyseur pour les deux formes vues précédemment :

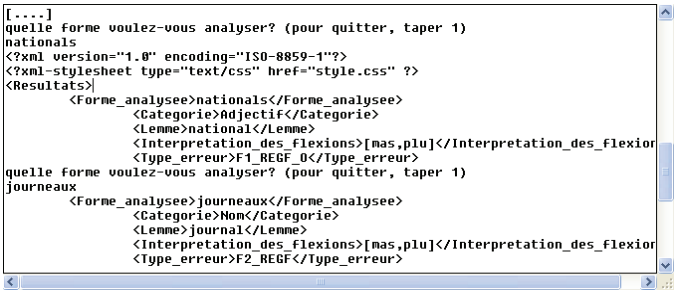


Figure 1 : Sorties de l'analyseur pour les formes *nationals* et *journeaux*.

Lors de l'analyse de *nationals*, l'algorithme découpe la forme en *national-* et *–s*. Il vérifie ensuite la présence de la base dans le dictionnaire. Grâce à cela, il récupère le modèle associé au lemme et compare les flexions et les régularisations effectuées lors du découpage par rapport aux flexions et régularisations autorisées (et obligatoires) du modèle. Il observe une incohérence sur une régularisation qui n'a pas été effectuée par l'apprenant pour former le masculin pluriel de *national*. Il génère un diagnostic en conséquence, puis tente un nouveau découpage (*nationa* + *ls*, etc.), sans succès.
Pour *journeaux*, aucun découpage ne permet d'extraire une base appartenant au dictionnaire. Il effectue donc une régularisation de forme (*-eaux* en *–als*), puis découpe la forme en *journal-* et *–s*. Les données associées au lemme ne correspondent pas aux données du

[7] http://www.prologia.fr/

découpage, au niveau de la régularisation qui est en fait une régularisation erronée. Il détecte ainsi une erreur et génère la rétroaction, comme montré ci-dessus (figure 1).

4 Conclusion et perspectives

Nous avons développé un module permettant la détection et l'analyse d'erreurs morphologiques en situation hors-contexte. Il donne des résultats tout à fait prometteurs quant à l'analyse des erreurs issues du corpus FRIDA (toutes les erreurs sont correctement analysées). Il est évident qu'il est nécessaire de tester ce module sur un corpus plus important que ce dernier, afin de valider notre démarche, et de compléter notre modélisation linguistique des erreurs morphologiques existantes.

Il serait intéressant d'intégrer ce module dans un système existant d'ALAO, afin de définir sa plus-value et la robustesse qu'il peut apporter à celui-ci. La prochaine étape de ce travail va ainsi consister à implémenter cette analyse au sein de la plateforme MIRTO (Antoniadis et al., 2005a) qui à l'heure actuelle se fonde sur une analyse à 4 niveaux (Kraif, 2005). Cette dernière n'est pas capable de générer des rétroactions sur des erreurs morphologiques. En effet, son analyse considère *chevals* comme une erreur d'orthographe, en trouvant une ressemblance graphique entre *chevals* et *cheval*, et ne peut extraire aucune information expliquant l'erreur de l'apprenant.
Ce module constitue une première pierre dans la réalisation d'un dispositif complet capable de détecter et interpréter les réponses d'apprenants. Il convient à présent d'élargir le type des erreurs à analyser. C'est dans cette optique que s'inscrivent mes travaux de thèse. Une première étape consistera à définir les types d'activités qui seront susceptibles d'être évaluées par des outils issus du TAL. Puis, nous établirons un corpus de productions d'apprenants via les activités retenues. Ces données permettront ensuite d'établir une typologie attestée des erreurs rencontrées. A partir de ces recherches, la conception d'un système TAL d'évaluation des productions et de génération de rétroactions pourra alors être menée.

Références

ÅGREN M. (2005). La morphologie du nombre dans le système verbal en français L2 écrit- L'accord de la 3ème personne du pluriel. Acquisition et production de la morphologie flexionnelle. *Actes du Festival de la morphologie*. Jonas Granfeldt et Suzanne Schylter Eds, Lund (Suède).

ANTONIADIS G., ECHINARD S., KRAIF O., LEBARBÉ T., PONTON C. (2005a). Modélisation de l'intégration des ressources TAL pour l'apprentissage des langues : la plateforme MIRTO. *Revue ALSIC*, 8.

ANTONIADIS G., FAIRON C., GRANGER S., MEDORI J., ZAMPA V. (2006). Quelles machines pour enseigner la langue. *TALN 2006*, Leuven.

ANTONIADIS G., PONTON C., ECHINARD S. (2005B). Le TAL au servie de l'évaluation automatique des fautes d'apprenants. Du vrai / faux à l'évaluation pédagogique. *Actes du colloque UNTELE*.

ASTOLFI J.P. (1997). *L'erreur, un outil pour enseigner*. Paris : Editions ESF.

BERRENDONNER A. (1990). *Grammaire pour un analyseur. Aspects morphologiques*. France : Les Cahiers du CRISS.

BOUILLON P., VANDOOVEREN F., DA SYLVA L., JACQMIN L., LEHMANN S., RUSSEL G., VIEGAS E. (1998). *Traitement automatique des langues naturelles*. Paris : Duculot.

CORDIER-GAUTHIER C., DION C. (2003). Correction et révision de l'écrit en français langue seconde : médiation humaine, médiation informatique. *Revue ALSIC*, 24,29-43.

FAYOL M. (2001). Compte-rendu de la conférence donnée par M. Fayol à l'IUFM de Poitou-Charentes, site des Deux-Sèvres à Niort, dans le cadre des Conférence de L'Ais : *Apprendre à utiliser l'orthographe*.

FUCHS C., DANLOS L., LACHERET-DUJOUR A., LUZATTI D., VICTORRI B. (1993). *Linguistique et Traitement Automatique des Langues*. France : Hachette.

GRANGER S., VANDEVENTER A., HAMEL M.J. (2001). *Analyse de corpus d'apprenants pour l'ELAO basé sur le TAL*. Paris : Hermès.

KRAIF O. (2005). Evaluation automatique de productions lexicales : une analyse à 4 niveaux. *Actes du colloque UNTELE*.

LALLICH-BOIDIN G., HENNERON G., PALERMATI R. (1990). *Analyse du français. Achèvement et implantation de l'analyseur morpho-syntaxique*. France : Les Cahiers du CRISS (Centre de Recherche en Informatique appliquée au Sciences Sociales.

LANCIEN T. (1997*). Le Multimédia (pp. 90-97)*. Clé International.

L'HAIRE S., VANDEVENTER A. (2003). Diagnostic d'erreurs dans le projet FreeText. *Revue ALSIC*, vol. 6, n°2, 21-37.

NIELSEN J. (1993). *Usability engyneering*. New York: Academic Press.

RÉZEAU J. (2004). *Médiatisation et médiation pédagogique dans un environnement multimédia. Le cas de l'apprentissage de l'anglais en Histoire de l'Art*. Thèse de doctorat de l'université Bordeaux 2 (pp. 356-367).

SILBERZTEIN M. (1993*). Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Paris : Masson.

WYATT D.H. (1988). *Applying pedagogical principles to CALL courseware development*, in Modem Media in Foreign Language Education : Theory and implementation. Illinois : National Textbook.

Repérage automatique de génériques dans les définitions terminographiques

Selja SEPPÄLÄ
Laboratoire de terminologie/TIM/ETI – Université de Genève
Bd du Pont-d'Arve 40, 1211 Genève 4
selja.seppala@eti.unige.ch

Résumé. Cet article présente une procédure de repérage et de balisage de l'élément générique de la définition terminographique exploitant les caractéristiques formelles du sous-langage définitoire. La procédure, qui comporte quatre étapes, constitue l'une des sous-tâches d'un analyseur (semi-)automatique de la structure conceptuelle des définitions terminographiques, destiné à faciliter l'annotation d'un corpus en vue de l'étude de régularités dans cette structure. La tâche décrite consiste à mettre au point un système d'annotation automatique basé sur le repérage d'indices morphosyntaxiques, sans recourir à d'autres ressources linguistiques informatisées.

Abstract. This article presents a procedure to locate and tag the generic elements of terminographic definitions, taking advantage of the formal characteristics of the definition sublanguage. This four step procedure is part of a larger (semi-)automatic parser of the conceptual structure of terminographic definitions, intended to ease the tagging of a corpus for studying conceptual regularities in definition structure. The method involves the development of an automatic tagging system, based on the identification of morphosyntactic boundary markers, which does not require the use of additional linguistic resources.

Mots-clés : définition terminographique, annotation automatique, repérage de frontière, indices morphosyntaxiques, sous-langage.

Keywords : terminographic definition, automatic tagging, boundary location, morphosyntactic markers, sublanguage.

1 Introduction

La présente communication s'inscrit dans le cadre d'une recherche en terminologie portant sur l'analyse de la structure conceptuelle de définitions terminographiques à travers l'étude d'un corpus de définitions annoté (Seppälä, 2004, 2005). L'un des objets de notre travail est la conception d'un analyseur (semi-)automatique de cette structure conceptuelle, qui facilite l'annotation du corpus en automatisant le plus grand nombre de tâches. Nous présentons ici l'une des étapes de l'annotation du corpus, à savoir la mise en place d'une procédure de

repérage et de balisage automatique des génériques (GEN) des définitions en compréhension, le générique étant l'élément qui rattache le concept défini à un concept plus général (souvent le *genre prochain*). La tâche décrite dans cet article est basée sur un traitement qui exploite les caractéristiques du sous-langage des définitions (Barnbrook, 2002). L'implémentation, en vue d'un balisage automatique des GEN des définitions, est faite à l'aide d'un programme Perl où les règles de repérage sont exprimées sous la forme d'expressions régulières.

Après une brève présentation de la définition terminographique, de sa structure générale et de l'élément générique (section 2), nous passons à l'identification du problème, afin de mieux cerner la tâche à accomplir (section 3). Nous abordons ensuite la description de l'expérience à proprement parler. Dans cette partie plus empirique, nous présentons tout d'abord les corpus utilisés (section 4), puis la méthode qui a été mise en place (section 5), pour terminer par une évaluation des performances de la procédure (section 6). En conclusion, nous évoquons les limites de cette méthode et quelques perspectives futures.

2 La définition terminographique

En terminographie, les définitions servent à décrire et distinguer les concepts d'un domaine (ou système conceptuel) spécialisé, plus ou moins clos, où les concepts entretiennent des relations généralement hiérarchiques. Elles consistent, dans la très grande majorité des cas, en une définition en compréhension[1]. Une définition en compréhension est composée, comme l'illustre l'exemple ci-dessous, d'une seule phrase où le GEN est suivi d'un ou de plusieurs spécifiques (SPE). Le SPE a pour fonction de préciser la portée du GEN en énumérant les traits ou propriétés du concept défini, tout en distinguant ce dernier des autres concepts du domaine auquel il appartient.

Ex : *peptidyl-transférase* = $_{GEN}$Enzyme $_{SPE1}$située dans le ribosome, $_{SPE2}$qui lie successivement un acide aminé supplémentaire à la chaîne polypeptidique en cours d'élongation.

La forme de ce type de définitions est relativement contrainte : elle se compose d'une seule phrase, où le GEN précède généralement le(s) SPE. Parfois, le GEN peut être précédé d'un adjectif – considéré ici comme SPE –, mais cela reste assez rare en français. À l'exception de ces cas de figure facilement identifiables, une définition respectant les conventions de rédaction terminographiques généralement admises (rappelées dans Seppälä, 2004, 2006) ne devrait pas être précédée d'un SPE en français.

Selon notre schéma d'analyse[2], chaque définition est segmentée en GEN et en SPE ; ces éléments sont ensuite annotés avec des étiquettes conceptuelles. La segmentation et l'annotation sont réalisées en XML, le but étant d'automatiser ces opérations au maximum. L'exemple suivant montre une version annotée de la fiche terminologique du concept *peptidyl-transférase*.

---

[1] 97 % d'un corpus étudié lors d'une précédente expérience (Seppälä, 2004, 2005), le reste étant des définitions en extension. Étant donné notre tâche, nous excluons de cette étude ces définitions, qui se composent d'une énumération d'espèces isonymes et n'ont donc pas de générique.

[2] Plusieurs niveaux d'analyse sont possibles selon la granularité voulue (Barnbrook, 2002). Pour plus de détails sur celui que nous avons adopté, voir (Seppälä, 2004).

Ex : <FICHE langue="FR">
<NI>25</NI>
<CM>biosynthèse des protéines</CM>
<VE>peptidyl-transférase</VE>
<DF><GEN relation_VE="GENRE"
classe_conceptuelle="NATUREL">Enzyme </GEN>
<SPE relation_GEN="SPATIAL">située dans le ribosome,</SPE>
<SPE relation_GEN="FONCTION">qui lie successivement un acide
aminé supplémentaire à la chaîne polypeptidique en cours
d'élongation.</SPE></DF>
</FICHE>[3]

L'élément générique de la définition, également appelé *incluant*, a pour fonction de rattacher le concept défini à un concept plus général. Sa forme présente un certain nombre de variations, néanmoins limitées et régulières. Le plus souvent il est constitué d'un seul mot[4]. Lorsque le GEN n'est pas un mot unique, il peut prendre la forme d'un terme complexe formé de deux mots ou plus (*Acide nucléique* ou *Mesures ou ouvrages*), d'un syntagme nominal (*Triplet de nucléotides*) ou encore d'un *faux incluant*[5] commençant par un marqueur relationnel (*Partie de…* ou *Ensemble de…*), dont la liste est relativement restreinte et qui permet d'annoter le GEN avec la relation conceptuelle (PARTIE, TOUT, GENRE ou GENRE PROCHAIN) qui l'unit au défini. Dans ce cas, c'est l'ensemble « marqueur + mot ou syntagme » qui constitue le GEN. Dans les deux cas, le GEN peut comporter une autre entrée du domaine (environ 50 % du corpus d'entraînement). En effet, les concepts définis en terminographie s'insèrent généralement dans un système conceptuel hiérarchisé, où le concept superordonné peut servir de générique à la définition du concept subordonné. Ceci se traduit, linguistiquement parlant, par la reprise dans le GEN, de l'un des termes ou synonymes (l'une des vedettes) désignant ce concept, comme le montre l'exemple suivant, où la vedette (*acide nucléique*) du concept superordonné devient le GEN de la définition du concept subordonné (*acide ribonucléique*).

Ex : ***acide nucléique*** = *Molécule constituée d'un enchaînement de nucléotides disposé le long d'un brin ou deux.*
acide ribonucléique = ***Acide nucléique*** *formé d'un seul brin et participant à toutes les étapes de la synthèse des protéines*.

Dans le cas des définitions d'adjectifs, peu fréquentes, c'est l'expression du type *Se dit de…* qui est considérée comme étant le GEN de la définition. Ainsi le GEN présente des régularités marquées qui contribuent à faire du discours définitoire un sous-langage spécifique (Barnbrook, 2002).

---

[3] NI = numéro d'identification unique ; CM = domaine ; VE = vedette (terme ou synonyme) ; DF = définition ; GEN = générique ; SPE = spécifique ; relation_VE = relation conceptuelle entre le GEN et la VE ; relation_GEN = relation conceptuelle entre le SPE et le GEN.

[4] Nous entendons par mot, une chaîne de caractères suivie d'un espace.

[5] On trouve le *faux incluant* dans cinq types de situation : lorsque la chose est définie par ses parties ; lorsqu'il y a définition de la chose transformée ; lorsque la chose est définie par sa cause ou sa conséquence ; lorsque l'incluant marque le rapport de la chose à l'unité ; et lorsqu'il y a faux incluant d'existence (Rey-Debove, 1971).

3 Identification de la tâche

Notre tâche est de repérer le générique de la définition et de le marquer, de part et d'autre, avec les balises XML suivantes : <GEN relation_VE="…">xxxxx</GEN>. Le contenu du GEN n'étant pas en cause, la question revient par conséquent à déterminer la frontière qui sépare le GEN du ou des SPE adjacent(s), selon qu'il est ou non précédé de cet élément. Dans la grande majorité des cas, le GEN se trouve en effet en début de définition terminographique ; il correspond même souvent au premier mot de la phrase (env. 70 % des GEN de notre corpus d'entraînement ; 97 % du corpus d'évaluation). On sait, par ailleurs, qu'un GEN est toujours suivi d'un SPE (en terminologie). Suivant le même constat que (Barnbrook, 2002), nous pouvons dire que la principale difficulté consiste donc à repérer sa limite à droite, plus exactement la frontière qui le sépare du premier SPE. Nous proposons d'exploiter les régularités du discours définitoire en identifiant ces frontières (début et fin de GEN) à l'aide de marqueurs morphosyntaxiques, lesquels peuvent ensuite être traduits en règles de repérage sous la forme d'expressions régulières.

La constitution d'une liste de marqueurs morphosyntaxiques nécessite une étude détaillée des principales caractéristiques des GEN et des SPE antéposés, ainsi que du début (limite à gauche) des SPE qui viennent immédiatement après le GEN. Pour ce faire, nous avons eu recours, d'une part, à la littérature (Iris, et al., 1988, Iso, 2000, L'homme, 2003, Rebeyrolle, 2000, etc.) et, d'autre part, à un corpus préalablement annoté à la main (voir section 4), afin d'y repérer les régularités susceptibles de servir d'indices de repérage. L'identification des marqueurs peut être réalisée manuellement et/ou automatiquement, par des méthodes d'apprentissage automatique. Dans la mesure où le discours définitoire constitue un sous-langage relativement contraint, présentant peu de variations, notamment au niveau des éléments morphosyntaxiques marquant la frontière entre les éléments de la définition (Barnbrook, 2002), nous avons opté pour un repérage manuel assisté d'un concordancier. L'identification d'indices peut également faire intervenir des traitements préalables (Vossen, et al., 1989), comme un étiquetage morphosyntaxique ou une lemmatisation ; la grande régularité du discours définitoire nous a là encore permis de nous en tenir à une méthode simple, basée sur les seuls éléments de surface, c'est-à-dire sur des chaînes de caractères.

Une fois les marqueurs identifiés et traduits en règles de repérage, il convient de mettre au point une procédure d'étiquetage qui n'applique qu'une seule règle par définition, afin d'éviter les erreurs d'annotation. Deux paramètres sont pris en compte :

- la hiérarchisation des étapes et des règles d'analyse de la plus spécifique à la plus générale, de façon à ce qu'elles entrent le moins possible en conflit entre elles, et
- la recherche systématique des vedettes (VE) de la base terminologique en début de définition, afin de pouvoir les inclure le cas échéant dans les GEN (étape appelée « GEN=VE »).

Selon ces constats, le repérage des SPE antéposés doit avoir préséance sur l'application des règles de repérage des GEN comprenant une VE, laquelle se doit de précéder la recherche de la fin du GEN lorsqu'on ne connaît pas son contenu, c'est-à-dire tous les autres cas. Pour terminer, il y a lieu d'évaluer la performance de ce système et d'envisager des pistes d'amélioration.

4 Présentation des corpus

Les expériences réalisées dans le cadre de cette étude portent sur deux corpus distincts : un corpus d'entraînement et un corpus d'évaluation. Le premier a servi à identifier les indices de repérage des GEN et les marqueurs pour chaque relation conceptuelle, ainsi qu'à tester, affiner et hiérarchiser les règles de repérage, de façon à obtenir la meilleure performance possible. Il s'agit d'un échantillon de 490 définitions en compréhension tirées de la *Banque de terminologie du canton de Berne* (Lingua-PC), où chaque définition a été préalablement segmentée et annotée à la main en GEN et en SPE[6]. Le second corpus a été utilisé pour tester la performance générale de la procédure d'annotation proposée. Il s'agit d'un ensemble de 92 définitions extraites d'un glossaire sur la *Terminologie de la biosynthèse des protéines chez les cellules eucaryotes* (Bourjault, 2005). Les deux corpus respectent les conventions de rédaction terminographiques et remplissant les critères de bonne formation.

5 Description du système

Dans cette partie, nous présentons l'architecture générale du système et une synthèse de ses principales étapes, en nous concentrant plus spécifiquement sur les difficultés qu'elles posent. Comme nous l'avons déjà souligné, les différentes étapes du traitement doivent être ordonnées pour éviter les conflits de règles. Nous distinguons quatre types de tâches à réaliser dans l'ordre suivant : repérage des SPE antéposés et marquage des balises de début de GEN (section 5.1) ; repérage d'éventuelles vedettes dans les génériques (GEN=VE) (section 5.2) ; repérage des fin de GEN (section 5.3) ; et finalement, marquage du premier mot des définitions non traitées au cours des étapes précédentes (section 5.4).

5.1 Repérage des spécifiques antéposés et marquage du début du GEN

Le repérage d'éventuels spécifiques antéposés et l'insertion des balises de début de GEN doivent précéder les autres étapes du traitement, au risque d'engendrer des problèmes de repérage des GEN. Une vedette utilisée en tant que générique pourrait, par exemple, ne pas être repérée, simplement parce qu'elle est précédée d'un SPE. Cette tâche consiste à assigner aux SPE antéposés les balises `<SPE>xxxxx</SPE><GEN>`. La dernière balise, qui est également ajoutée au début de toutes les autres définitions, marque le début du GEN et est nécessaire pour que les règles suivantes puissent être uniformément appliquées à l'ensemble des définitions. Les SPE antéposés peuvent être de deux types – adjectifs et indication de domaine –, dont les caractéristiques formelles susceptibles de servir d'indice à leur repérage sont aisément identifiables et peu nombreuses.

1. Adjectifs : La nature non indexicale, objective et factuelle des définitions terminographiques implique qu'en français, la liste des adjectifs antéposables pouvant servir de marqueur de SPE antéposé est relativement réduite. Le corpus d'entraînement, par exemple, ne fait état que d'adjectifs numéraux ordinaux (*première*, *deuxième*, etc.) et de quelques adjectifs qualificatifs (*grand*, *petit*, etc.) ou de leurs superlatifs. Formellement parlant, ces adjectifs donnent lieu à deux types de SPE antéposés :

[6] Pour une description plus détaillée du corpus et de l'annotation conceptuelle, voir (Seppälä, 2004, 2005).

- Le premier consiste en un adjectif qui peut être précédé d'un superlatif absolu et qui est immédiatement suivi du GEN.
  Ex : <SPE> (superlatif absolu)[7] + adj. </SPE> + <GEN>
  ⇒ *très grand [vaisseau]* ou *première [phase]*

- Le second consiste en un adjectif, superlatif relatif ou numéral ordinal, précédé d'un article défini et éventuellement d'un comparatif, et suivi d'un article indéfini et éventuellement d'un nombre.
  Ex : <SPE> art. défini + (comparatif) + adj. + art. indéfini + (nombre) </SPE> + <GEN>
  ⇒ *le plus grand des [singes]* ou *la première des trois [étapes]*

2. Indication de domaine : Dans les bases de données terminologiques, l'indication du domaine et des sous-domaines se fait dans un champ d'indexation propre. Or, ce type d'indication apparaît parfois en début de définition, souvent pour restreindre la portée du concept défini à un sous-domaine plus spécifique. Une définition bien formée ne devrait pas inclure ce type d'élément, mais un analyseur automatique de définitions doit néanmoins prévoir ces cas de figure. La forme de cette indication est, elle aussi, très contrainte : elle commence par une préposition, suivie d'un ou de plusieurs mots, et se termine par une virgule.
   Ex : <SPE> En | Dans | Sur + mot (s) quelconque(s) + virgule </SPE> + <GEN>
   ⇒ *En droit constitutionnel,…* ou *Sur une locomotive,…*

5.2 Repérage des génériques incluant une vedette

La deuxième étape consiste à vérifier si le GEN reprend l'un des termes du domaine. Pour ce faire, toutes les vedettes de la base sont extraites dans une liste (18080 termes dans le corpus d'entraînement), qui doit être triée par ordre alphabétique, du terme le plus long au plus court, afin d'éviter que certains éléments des termes complexes ne soient exclus du GEN. L'exemple suivant montre la séquence à respecter pour les termes complexes dérivés d'une même base lexicale : *traitement annuel déterminant* → *traitement annuel* → *traitement*. Le programme vérifie ainsi pour chaque VE si elle apparaît en début de définition ou à la suite d'une expression relationnelle. Si c'est le cas, l'élément correspondant est marqué comme GEN et se voit attribuer la relation GENRE PROCHAIN ou celle qui correspond au marqueur de relation qui le précède ; sinon, la définition est remise dans le circuit pour être traitée par l'une des règles présentées dans la section suivante. Le développement de cette tâche présente deux types de difficultés.

1. Deux VE pour un même GEN : Il arrive que le programme repère une VE plus longue alors que la plus courte existe aussi dans la liste et que c'est celle-là qu'il aurait convenu de marquer : la liste contient par exemple les trois VE *crédit de paiement*, *crédit d'engagement* et *crédit*. Si la définition commence par *Crédit de paiement ou d'engagement…*, le programme repère tout d'abord *crédit de paiement* et le marque comme GEN. Or, il serait dans ce cas plus adéquat, soit de ne retenir que *crédit* et de considérer les deux éléments suivants comme un SPE précisant la nature de ce crédit, soit de marquer les deux comme composant un seul GEN.

---

[7] Les parenthèses marquent les éléments optionnels ; le « | » sépare différentes variantes ; et les crochets indiquent le GEN.

2. Traitement des pluriels : Les termes en vedette sont généralement sous la forme canonique, donc au singulier. La liste des VE prise en entrée du programme comporte donc principalement cette forme. Dans un GEN, ces termes peuvent en revanche apparaître au pluriel, ce qui n'a pas d'incidence lorsque, pour des questions d'usage (certains termes ne s'utilisent qu'au pluriel), ils figurent déjà au pluriel dans la base. C'est en revanche plus problématique lorsqu'il faut repérer la forme plurielle d'une VE au singulier, en particulier dans les cas de pluriels irréguliers, mais surtout dans ceux des mots composés. Notons toutefois que le taux d'erreurs dues à ce type de cas devrait rester très faible et qu'il peut être aisément réduit dans la mesure où, dans des définitions bien formées, seuls les termes qui suivent une expression relationnelle tendent à apparaître au pluriel, et que ces cas sont pris en charge par les règles plus générales de la troisième étape.
   Lors du développement, nous avons tout de même constaté que le fait de prévoir le repérage de VE suivies de la marque du pluriel la plus générale, à savoir un « s », permet d'améliorer la précision (VE = *dépense d'investissement* → GEN = *dépense d'investissements*). Seule réserve, mais plutôt relative au fond : il arrive que le GEN marqué et le pluriel de la VE reconnue dans la liste soient homographes, auquel cas ce n'est pas le véritable terme de la base qui est repéré, mais un autre terme au singulier (par exemple, VE : sg. *fond* → pl. *fonds* et GEN : *fonds*, mais ce dernier n'est pas le pluriel de *fond*, il s'agit d'un autre terme au singulier). Cependant, l'élément ainsi étiqueté reste un vrai GEN et le résultat est considéré comme acceptable, étant donné que la tâche est ici de trouver la frontière des GEN.

5.3 Repérage des fins de générique

La troisième étape consiste à soumettre toutes les définitions traitées lors de la première étape, mais ignorées lors de la deuxième, aux règles de repérage des marqueurs de frontière de fin de GEN. Nous avons vu que les définitions terminographiques présentent beaucoup de régularités, notamment au niveau des frontières autour desquelles s'articulent les GEN et les SPE. D'après (Barnbrook, 2002), un analyseur peut identifier ces limites en utilisant une combinaison de trois éléments : une règle générale identifiant les participes présents et passés réguliers ; une liste de moins de 100 mots comportant des membres des classes fermées (*selon*, *qui*, *dans*, etc.) et des participes passés irréguliers ; ainsi qu'une liste d'exclusion comprenant les mots susceptibles d'être mal traités par la règle de repérage générale. Bien que l'analyse que nous envisageons ne soit pas tout à fait la même, la présence des mêmes types de régularités a néanmoins été vérifiée sur notre corpus d'entraînement. Suivant ce constat, nous avons établi des règles correspondant à ces différents cas de figure, en prenant en compte toutes les réalisations possibles d'un marqueur. La marque du participe présent en *–ant* doit par exemple être déclinée en *–ante*, *–ants* ou *–antes*.

Les règles visant à contraindre le repérage de la fin du GEN exploitent donc les caractéristiques des contextes droit et gauche du GEN, et gauche du SPE qui le suit, et tiennent compte du nombre de mots apparaissant généralement entre les deux. Nous distinguons deux types de règles : des règles de fin spécifiques et des règles de fin générales.

1. Les règles spécifiques concernent les cas de faux incluants et exploitent à la fois les expressions relationnelles en début de GEN (*Type de…*, *Ensemble de…*, *Partie de…*, etc.), dont le nombre est restreint et qui permettent d'associer une relation spécifique au GEN (GENRE, TOUT ou PARTIE), et la ponctuation à la fin du GEN. (Dans les cas de GEN=VE, ce sont les VE qui constituent la limite à droite du GEN.)

2. Les règles générales concernent tous les autres cas et se basent sur trois types de marqueurs :

- sur la terminaison du premier mot après le GEN ($_{fin}MOT_{GEN+1}$), sachant qu'il s'agit généralement d'un participe présent ou passé ;
- sur le premier mot qui suit le GEN (MOT_{GEN+1}). Dans ce cas, les règles (lexicales) intègrent un certain nombre de mots entiers, tels que des conjonctions (*que*, *quand*, etc.), des pronoms relatifs (*qui*, *tel*, etc.) ou des prépositions (*par*, *dans*, etc.).
- Soit enfin sur la ponctuation qui suit le GEN, à savoir la virgule ou la parenthèse.

La principale difficulté liée au développement de ces règles est de les hiérarchiser en sorte qu'une règle plus générale ne marque pas des GEN qui devraient être traités par une règle plus spécifique. L'exemple suivant illustre un conflit de règles où l'étiquetage est faux si la règle lexicale plaçant le GEN avant *par*, précède la règle qui place la frontière avant un mot qui se termine par *–é*. Si on inverse l'ordre, le résultat est juste.

FAUX : *Gain brut réalisé</GEN> par…*
VRAI : *Gain brut</GEN> réalisé par…*

Une autre difficulté est liée à la préposition *de*. Ce marqueur apparaît aussi bien dans les mots composés à l'intérieur du GEN, qu'après un GEN (en début de SPE). La solution adoptée jusqu'à présent est de considérer que la préposition marque le début d'un SPE et que le GEN est donc composé d'un seul mot. Cette solution est en effet la plus probable étant donné que près de 70 % des GEN du corpus d'entraînement et 60 % du corpus d'évaluation sont des mots simples. Il semblerait aussi que les cas ambigus soient souvent mieux traités par les autres règles générales de type $_{fin}MOT_{GEN+1}$. La règle du marqueur *de* doit donc être classée parmi les dernières.

Le classement des règles est réalisé manuellement, en fonction de la fréquence d'application et de la performance de chaque règle prise individuellement, ainsi que des observations faites lors des tests et des ajustements successifs. L'ordre de classement général est le suivant :

1. Les règles spécifiques sont appliquées en premier (y compris dans l'étape GEN=VE), étant donné qu'elles sont contraintes par le début du GEN. Comme elles n'entrent pas en conflit entre elles, leur ordre interne n'a pas d'importance.
2. Suivent les règles lexicales permettant de repérer des mots MOT_{GEN+1} (*auquel*, *dont*, *lors*, *où*, etc.) qui précédent les participes.
3. Viennent ensuite les règles $_{fin}MOT_{GEN+1}$, qui précisent la fin du premier mot après le GEN. Celles-ci non plus n'entrent pas en conflit entre elles, dans la mesure où les formes des patrons recherchés sont bien distinctes. Il est donc possible des les placer sans ordre spécifique dans un même bloc précédant
4. une seconde règle lexicale qui comporte des prépositions (*dans*, *entre*, *par*, *pour*, *sur*) ne pouvant être placées avant les règles recherchant des participes.
5. Viennent finalement les règles les plus « bruyantes », qui tendent à provoquer beaucoup d'erreurs lorsqu'elles sont placées avant les autres (préposition *de* et *ponctuation*).

Dans tous les cas, il conviendrait de tester et d'ajuster l'ensemble des règles sur un corpus plus large, et de les classer ensuite statistiquement de façon à obtenir la meilleure performance possible.

5.4 Étiquetage des définitions non traitées

La quatrième et dernière étape consiste à soumettre au même traitement toutes les définitions dont les GEN n'auraient pas encore été balisés. Il s'agit de leur appliquer la règle la plus générale (règle par défaut) qui marque systématiquement le premier mot de la définition comme étant le GEN. L'application de cette seule règle à la troisième étape (une fois les GEN=VE exclus et en ignorant les autres règles) montre que 72,7 % des définitions du corpus d'entraînement auraient été étiquetées convenablement. L'ensemble des règles de repérage de fin de GEN (celles de la 3e étape) visent donc à couvrir les 27,3 % de cas d'échec de la règle par défaut (celle qui marque systématiquement le 1er mot). Ce constat réduit en fait considérablement le nombre d'exemples du corpus d'entraînement permettant d'apprécier toute la gamme des variations possibles de la forme du GEN, mais cela n'empêche en rien l'étude des marqueurs post-GEN qui indiquent la frontière à droite de l'élément.

6 Évaluation de la procédure

L'évaluation réalisée sur le second corpus montre que la performance générale de ce système est relativement bonne : 88 % des GEN sont correctement annotés, ce qui constitue une très nette amélioration par rapport à une annotation par défaut du 1er mot (60 %) de chaque définition. Cette amélioration de la performance est due à la prise en compte des VE (57 % des GEN) et des expressions relationnelles (8 %), ainsi que, dans une moindre mesure, au repérage des SPE antéposés (quelque 3 %). La plupart des erreurs (7/11 erreurs) sont dues aux règles générales, dont la performance pourrait être améliorée en les affinant davantage ou en les hiérarchisant différemment. Un apprentissage automatique de ces règles pourrait également être envisagé. Les autres erreurs (4/11) sont dues au non repérage d'un faux incluant, qui s'explique généralement par l'absence d'un des éléments clés du marqueur, comme les mots soulignés dans ces GEN : *Branche de la biologie* ou *Constituant d'une cellule*. Ces erreurs peuvent être corrigées en insérant le mot en question dans la règle, au risque cependant de la rendre plus ambigüe, notamment si elle est appliquée à un autre domaine. Ce serait sans doute le cas du mot *région* du GEN *Région de l'ADN*. Pour voir dans quelle mesure les marqueurs sont spécifiques à un domaine et vérifier ces résultats, il conviendrait de poursuivre l'évaluation sur un plus grand nombre de données, provenant de domaines variés.

7 Conclusion

Dans cet article, nous avons présenté un procédé de repérage et d'étiquetage des éléments génériques des définitions terminographiques, qui exploite les régularités formelles du sous-langage définitoire, afin de créer des règles de repérage les plus précises possibles de la frontière entre le générique et le(s) spécifique(s). Il ne nécessite aucun prétraitement morphosyntaxique ni recours à des ressources de TALN spécifiques, mais exige que les différentes étapes du traitement et les règles, susceptibles d'entrer en conflit, soient hiérarchisées. Après une présentation des quatre étapes de traitement des définitions et des difficultés inhérentes à leur application, nous avons montré que le système permet d'obtenir de bonnes performances globales. Cette performance est principalement obtenue grâce aux règles qui repèrent les spécifiques antéposés, les termes du domaine et les expressions relationnelles. Les résultats devraient cependant être consolidés sur des corpus plus grands, couvrant différents domaines, et des améliorations pourraient être étudiées en ayant recours à

des procédés statistiques ou d'apprentissage automatique. Finalement, si l'architecture de ce système semble transposable à d'autres langues, l'exploitation des variations morphosyntaxiques suppose toutefois que des règles de repérage spécifiques soient créées pour chaque langue, ce qui restreint considérablement leur réutilisabilité. Des techniques d'analyse indépendantes des langues basées sur l'apprentissage automatique des règles devraient permettre de pallier cette limite. Nous envisageons à l'avenir d'explorer ces pistes statistiques afin d'optimiser les performances du système. Nous étudierons également la possibilité d'étendre ce type de procédé à l'annotation des spécifiques, l'objectif final étant de disposer d'un analyseur qui facilite l'annotation de corpus de définitions terminographiques, en vue d'en étudier les éventuelles régularités dans la structure conceptuelle. Ce type d'analyseur pourrait finalement s'avérer intéressant pour enrichir les options de recherche dans les bases de données terminologiques.

Références

BARNBROOK, G. (2002). *Defining Language : A local grammar of definition sentences*. Amsterdam, Philadelphia: John Benjamins.

BOURJAULT, A. (2005). *Terminologie de la biosynthèse des protéines chez les cellules eucaryotes : anglais-français*. Université de Genève, École de traduction et d'interprétation.

IRIS, M. A., et al. (1988). Problems of the part-whole relation. *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, 261-287.

ISO (2000). *Travaux terminologiques : principes et méthodes (ISO 704)*. Genève: ISO.

L'HOMME, M.-C. (2003). Indices de relations conceptuelles dans les définitions terminologiques. Application au domaine de l'informatique. Actes de *I Jornada Internacional sobre la Investigación en Terminología y Conocimiento Especializado*, 44-50.

LINGUA-PC. *Banque de terminologie du canton de Berne*. Chancellerie d'État du Canton de Berne.

REBEYROLLE, J. (2000). Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. Actes d'*IC'2000*.

REY-DEBOVE, J. (1971). *Étude linguistique et sémiotique des dictionnaires français contemporains*. The Hague, Paris: Mouton.

SEPPÄLÄ, S. (2004). *Composition et formalisation conceptuelles de la définition terminographique*. Université de Genève, École de traduction et d'interprétation.

SEPPÄLÄ, S. (2005). Structure des définitions terminographiques : une étude préliminaire. Actes de *Terminologie et Intelligence Artificielle, TIA'05*, 19-29.

SEPPÄLÄ, S. (2006). Semi-Automatic Checking of Terminographic Definitions. Actes de *TermEval Workshop - LREC 2006*, 22-27.

VOSSEN, P., et al. (1989). Meaning and structure in dictionary definitions. *Computational Lexicography for Natural Language Processing*, 171-192.

Extraction de paraphrases désambiguïsées à partir d'un corpus d'articles encyclopédiques alignés automatiquement

François-Régis CHAUMARTIN
Lattice/Talana – Université Paris 7
30 rue du château des rentiers, 75013 Paris
fchaumartin@linguist.jussieu.fr, frc@proxem.com

Résumé. Nous décrivons ici comment enrichir automatiquement WordNet en y important des articles encyclopédiques. Ce processus permet de créer des nouvelles entrées, en les rattachant au bon hyperonyme. Par ailleurs, les entrées préexistantes de WordNet peuvent être enrichies de descriptions complémentaires. La répétition de ce processus sur plusieurs encyclopédies permet de constituer un corpus d'articles comparables. On peut ensuite extraire automatiquement des paraphrases à partir des couples d'articles ainsi créés. Grâce à l'application d'une mesure de similarité, utilisant la hiérarchie de verbes de WordNet, les constituants de ces paraphrases peuvent être désambiguïsés.

Abstract. We describe here how to automatically import encyclopedic articles into WordNet. This process makes it possible to create new entries, attached to their appropriate hypernym. In addition, the preexisting entries of WordNet can get enriched with complementary descriptions. Reiterating this process on several encyclopedias makes it possible to constitute a corpus of comparable articles; we can then automatically extract paraphrases from the couples of articles that have been created. The paraphrases components can finally be disambiguated, by means of a similarity measure (using the verbs WordNet hierarchy).

Mots-clés : extraction de paraphrases, fusion d'articles, mesure de similarité, distance sémantique, identification d'hyperonyme, WordNet, Wikipedia, entités nommées, analyse syntaxique, désambiguïsation lexicale, cadres de sous-catégorisation, apprentissage.

Keywords: paraphrases extraction, articles merging, similarity measure, semantic distance, hypernym identification, WordNet, Wikipedia, named entities, syntactic analysis, word sense disambiguation, syntactic frames, unsupervised learning.

1 Introduction

1.1 Architecture d'ensemble

Nous souhaitons disposer d'une correspondance directe entre les articles d'une encyclopédie et les entrées d'un lexique sémantique de référence. Deux cas de figure se rencontrent alors ;

quand une entrée de lexique correspond déjà à un article, nous établissons la correspondance entre les deux ; sinon, nous enrichissons le lexique, en créant une nouvelle entrée et en la rattachant (via une relation d'hyperonymie/hyponymie) au meilleur « ancêtre » existant.

En réitérant ce processus sur plusieurs encyclopédies, nous obtenons un corpus monolingue de paires d'articles traitant d'un même sujet, propice à la découverte de paraphrases. Nous pouvons alors déterminer, par exemple, que « *la rivière Alabama serpente jusqu'à Selma* » est une paraphrase de « *la rivière Alabama coule vers Selma* ». Nous représentons les paraphrases sous forme de triplets (sujet, verbe, complément). La désambiguïsation des entités nommées permet d'établir que « RIVIÈRE$_{\#1}$ serpente (préposition) VILLE$_{\#1}$ » est une paraphrase de « RIVIÈRE$_{\#1}$ coule (préposition) VILLE$_{\#1}$ ». (L'indice $_{\#i}$ indique le sens du mot dans le lexique.) L'utilisation d'une mesure de similarité entre les deux verbes permet enfin de déterminer les sens de « serpenter » et « couler » dans le contexte. Nous obtenons, au final, l'équivalence entre deux cadres de sous-catégorisation, dont les éléments sont désambiguïsés par rapport au lexique : SERPENTER$_{\#1}$ (RIVIÈRE$_{\#1}$, VILLE$_{\#1}$) ~ COULER$_{\#2}$ (RIVIÈRE$_{\#1}$, VILLE$_{\#1}$).

Ces opérations constituent les deux premières étapes du projet ISIDORE[1], qui vise à extraire des connaissances d'une encyclopédie en langue anglaise. Pour faciliter la lecture, les exemples cités ici ont été traduits en français.

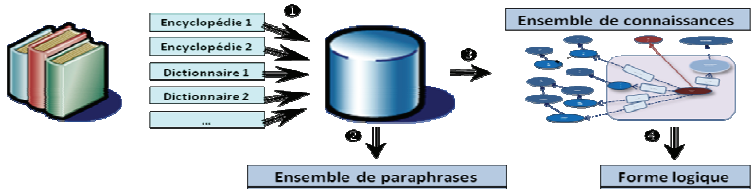


Figure 1 : architecture d'ensemble du projet ISIDORE

1.2 Lexique de référence

Notre lexique de référence est WordNet (Miller, 1995) version 2.1. Ce projet, mené depuis 1985 à Princeton, offre un réseau sémantique très complet de la langue anglaise. S'il n'est pas exempt de critiques (granularité très fine, absence de relations paradigmatiques…), WordNet n'en reste pas moins l'une des ressources de TAL[2] les plus populaires.

Les nœuds sont constitués par des ensembles de synonymes (ou *synsets*), correspondant au sens d'un ou plusieurs lemmes. Un synset est défini d'une façon différentielle par les relations qu'il entretient avec les sens voisins. Par exemple, des relations d'hyperonymie et d'hyponymie relient les « ancêtres » des noms et des verbes avec leurs « spécialisations ». La version 2.1 a de plus introduit la notion d' « instance hyponyme », qui désigne une instance

[1] St-Isidore (560-636), patron des informaticiens, fut l'auteur des *Etymologies*, une encyclopédie en 20 livres.

[2] WordNet est téléchargeable sur http://wordnet.princeton.edu.

(typiquement une entité nommée) d'un synset, et non une sous-classe. Ainsi, le nom TOUR$_{\#1}$ a SILO$_{\#1}$, MINARET$_{\#1}$, PHARE$_{\#1}$… pour hyponymes, et TOUR EIFFEL $_{\#1}$ comme instance hyponyme.

2 Importation d'articles encyclopédiques dans WordNet

L'encyclopédie en ligne *Wikipedia* possède une vingtaine d'articles dont le titre contient (au moins partiellement) « *Abraham Lincoln* » :

1. « *Abraham Lincoln* » : l'homme politique, 16ème Président des Etats-Unis.
2. « *Abraham Lincoln assassination* » : l'assassinat de l'homme politique.
3. « *Abraham Lincoln (Pullman car)* » : le plus ancien wagon de passagers des Etats-Unis.
4. Sans oublier deux films biographiques, trois lieux géographiques, plusieurs écoles, deux vaisseaux militaires… également nommés en mémoire de l'homme politique.

Nous constatons donc qu'une similarité entre le titre d'un article et un lemme (ou groupe de mots) désignant un synset de WordNet ne suffit pas à déduire qu'ils traitent du même sujet.

Nous cherchons à identifier le (ou les) synset de WordNet auquel un article se rattache. Pour ce faire, nous commençons par extraire de WordNet les « synsets candidats » pouvant correspondre au titre de l'article. Cette étape ne pose pas de difficulté particulière. Pour les personnes, par exemple, chaque article possède un ou plusieurs titres normalisés (de la forme « Prénom Nom » ou « Nom, Prénom »). Il suffit de rechercher les synsets correspondants dans WordNet. Pour un nom commun, il est nécessaire de tenir compte d'éventuelles variantes morphologiques et de retrouver la forme de base du mot. Nous appliquons alors un ensemble d'heuristiques[3] pour retenir le meilleur candidat. S'il n'en existe pas, nous commençons par chercher le synset correspondant le mieux au thème de l'article (décrit-il une rivière, un président… ?) Ensuite, nous créons un nouveau synset, rattaché (en tant qu'hyponyme ou instance hyponyme) au synset du thème de l'article.

Dans l'univers du traitement automatisé des encyclopédies, la *Wikipedia* pose un problème particulier. Pouvant être modifiée par tout internaute, elle voit depuis plusieurs années une progression exponentielle de son nombre d'entrées[4] : certain articles ne sont que des biographies auto-promotionnelles, d'autres des comptes-rendus de films ou de jeux vidéo… Notre choix est de ne retenir que les entrées correspondant à un consensus en termes de connaissances encyclopédiques. Nous travaillons donc sur un sous-ensemble des articles de la *Wikipedia* recoupant (sur la base du titre) ceux d'une autre encyclopédie de référence.

[3] (Carré, Degremont, Gross, Pierrel, Sabah, 1991) définit (p. 48) une heuristique comme « une règle qu'on a intérêt à utiliser en général, parce qu'on sait qu'elle conduit souvent à la solution, bien qu'on n'ait aucune certitude sur sa validité dans tous les cas ».

[4] 1 539 908 fin 2006 ; 874 359 fin 2005 ; 414 023 fin 2004 ; 188 538 fin 2003 ; 95 735 fin 2002.

2.1 Autre projet similaire

(Ruiz-Casado, Alfonseca, Castells, 2005) présentent l'implémentation d'un algorithme rapide permettant de réaliser la correspondance entre un article de la *Simple Wikipedia*[5] et le synset correspondant de WordNet[6]. Si aucun synset n'a de lemme en commun avec le titre de l'article, ce dernier est ignoré. Si un seul synset de WordNet a un lemme égal au titre, l'article y est lié sans autre analyse. En cas d'ambiguïté, l'article fait l'objet d'un étiquetage morphosyntaxique (après un filtrage des marqueurs syntaxiques spécifiques à la *Wikipedia*), pour ne conserver que les noms, verbes et adjectifs. Le système analyse les définitions de WordNet, et construit pour chacune d'entre elles un vecteur booléen (contenant « 1 » pour chaque terme en commun avec l'article et « 0 » pour chaque mot en disjonction). L'algorithme calcule alors une mesure de type cosinus entre les vecteurs, et retient le meilleur article, au sens de cette mesure de similarité.

2.2 Heuristiques utilisées dans notre approche

Notre approche améliore celle présentée ci-dessus, avec deux différences. D'une part, nous avons ajouté plusieurs heuristiques, afin d'augmenter la précision. D'autre part, nous appliquons ces heuristiques même dans le cas où un seul synset de WordNet a un lemme égal au titre de l'article. Comme nous l'avons vu, la *Wikipedia* ne contient pas moins de vingt articles sur « *Abraham Lincoln* » ; cette décision permet d'éviter des appariements erronés.

Les heuristiques utilisées sont indépendantes les unes des autres ; elles peuvent donc être appliquées dans n'importe quel ordre. Au départ, tous les synsets candidats partent avec un même indice de confiance, qui est modifié durant l'application des heuristiques. Après cette étape, les synsets candidats qui disposent d'un poids manifestement trop faible pour correspondre à l'article sont supprimés de la liste. Dans notre cas, nous avons déterminé expérimentalement un poids minimal de 0,6. Ensuite, on conserve les synsets dont l'indice de confiance vaut au moins 40% de celui du synset le mieux classé. Ceci permet de supprimer les synsets non significatifs.

2.2.1 Distance vectorielle sur les mots

Cette heuristique est identique à celle décrite dans (Ruiz-Casado, Alfonseca, Castells, 2005).

2.2.2 Comparaisons des contextes (domaines implicites et noms propres)

Nous extrayons du texte les domaines (« biologie», « sport »…) éventuellement associés à chaque mot[7], ainsi que les noms propres. Nous comparons la liste d'éléments extraits de l'article avec celle de chaque synset candidat, également à l'aide d'une mesure vectorielle.

---

[5] Une version en anglais simplifié de la Wikipedia (http://simple.wikipedia.org).

[6] Les auteurs revendiquent une précision de 91,11% (83.89% sur les mots polysémiques).

[7] WordNet associe parfois explicitement un domaine (baseball, géologie, mathématiques…) à un synset. Dans cette étape, nous comptons les domaines associés à chaque sens possible d'un mot du contenu de l'article.

2.2.3 Comparaison des domaines cités explicitement dans le texte

Cette heuristique recherche, dans une définition, des patrons de la forme « *en mathématiques* », « *utilisé en géologie* »… à l'aide d'expressions régulières. Si un patron de ce type est repéré, son domaine d'application est extrait (« mathématiques » ou « géologie » par exemple). Si le synset candidat (ou l'un de ses hyperonymes) appartient à ce domaine, son indice de confiance est augmenté.

2.3 Comparaison des hyperonymes

Cette heuristique a pour but de déterminer l'hyperonyme du sujet de l'article, en étudiant sa définition. En voici quelques exemples, où les hyperonymes sont soulignés :

- **Abraham Lincoln** : *16ème Président des Etats-Unis*.
- **Australie** : *un pays et le continent le plus petit.*
- **chat** : *mammifère félin ayant une épaisse fourrure douce et incapable de rugir*.

Le ou les hyperonymes du sujet de l'article sont comparés aux hyperonymes des synsets candidats. S'ils sont suffisamment proches (au sens d'une mesure de similarité), l'indice de confiance est fortement augmenté. Cette heuristique est essentielle en termes d'amélioration de la précision de l'appariement ; c'est pourquoi elle est détaillée ici.

2.3.1 Analyse syntaxique de la définition

Notre but est d'extraire l'hyperonyme d'une définition. Prenons l'exemple précédent du « chat » ; notre but est d'extraire « *mammifère* » (ou éventuellement « *mammifère félin* », si ce terme existe dans le lexique de référence)[8].

Nous effectuons pour cela une analyse syntaxique en profondeur de la définition, en utilisant le *Stanford Parser*[9] (Manning, Klein, 2002). Cet analyseur statistique fournit une sortie sous forme de dépendances syntaxiques.

Nous supposons que l'hyperonyme se situe dans la 1ère phrase de l'article, qui tient le plus souvent lieu de définition ; nous ne traitons donc que celle-ci. Comme une définition se résume souvent à un groupe nominal, il convient de la modifier pour la rendre « grammaticalement correcte ». Notre expérience montre que c'est indispensable dans le cas d'un analyseur basé sur des règles comme le *Link Grammar Parser* (Sleator, Temperley, 1991) et souhaitable dans le cas d'un analyseur statistique tel que le *Stanford Parser*. La première passe consiste donc en un étiquetage morphosyntaxique de la définition ; ensuite, en fonction de la partie du discours (adjectif, nom, verbe, etc.) du premier mot, l'algorithme préfixe éventuellement la définition par « *c'est* » ou « *c'est un* ».

---

[8] Si l'hyperonyme est qualifié par un adjectif ou un complément de nom, l'algorithme teste l'existence d'un synset constitué par l'expression complète, de façon à être le plus précis possible.

[9] Composant Java téléchargeable sur http://nlp.stanford.edu/downloads/lex-parser.shtml.

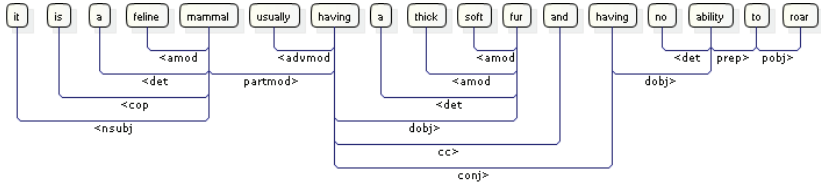


Figure 2 : Analyse syntaxique de la définition (en anglais) du nom « *chat* »

2.3.2 Recherche de l'hyperonyme

L'analyse syntaxique de la définition est alors disponible sous forme d'un graphe de dépendances. Nous le transformons en clauses Prolog, à partir desquelles nous pouvons identifier des schémas (Chaumartin, 2006).

Le processus tient compte des conjonctions de coordination, afin d'extraire correctement les hyperonymes multiples comme dans « *l'Australie est un pays et le continent le plus petit* ». Dans une construction comme « *une espèce de…* » ou « *un membre du groupe de…* », nous remontons d'une façon récursive le long des constituants de l'amas nominal, en passant au constituant imbriqué suivant.

2.3.3 Création de nouveaux synsets

Si aucun synset de WordNet ne correspond à l'article considéré, on en crée un nouveau, dont la définition sera la première phrase de l'article. Ensuite on le relie au synset représentant l'hyperonyme de l'article étudié. On est confronté ici à une problématique de désambiguïsation lexicale, pour identifier le sens correct. Par exemple, si l'hyperonyme est « *empereur* », il faut choisir entre les sens « *dirigeant mâle d'un empire* », « *raisin rouge de Californie* » ou « *grand papillon richement coloré* ».

Les hyponymes du meilleur ancêtre se situent au même niveau que le sujet de l'article dans la hiérarchie de WordNet. Nous cherchons donc des points communs entre l'article et ses « cousins » potentiels. Nous commençons par relever les similarités au niveau du vocabulaire employé entre l'article et chacun des hyponymes de ses ancêtres possibles ; en effet, des articles ayant le même hyperonyme ont une forte probabilité de traiter de sujets voisins, et donc de partager un champ lexical.

Pour finir, nous appliquons deux heuristiques supplémentaires. Tout hyperonyme candidat d'une entité nommée (personne, lieu, etc.) voit son indice de confiance augmenté si :

- Il en découle des relations de type « instance hyponyme ».
- Il hérite d'un groupe social (« *entreprise* », « *organisation* », « *mouvement* »…).

2.4 Résultats obtenus pour l'appariement d'articles

La version de mars 2006 de la *Wikipedia* en anglais (1 005 682 articles) a été filtrée pour retenir 15 847 articles, dont le titre était également présent dans une autre encyclopédie de

référence. Ces articles ont été appariés automatiquement sur WordNet. Pour évaluer la précision de l'appariement, nous avons examiné manuellement le résultat sur 800 articles :

- 505 ont été associés à un synset existant déjà dans WordNet ; l'appariement a été fait correctement dans 465 cas (soit une précision de 92%).
- 295 nouveaux synset ont été crées ; l'hyperonyme a été correctement identifié dans 251 cas (soit une précision de 85%).

2.5 Bilan : constitution d'un corpus monolingue d'articles comparables

En répétant le processus précédent sur plusieurs sources encyclopédiques, nous pouvons rattacher plusieurs articles à un même synset, et obtenir un corpus d'articles comparables.

Wikipedia	Encyclopédie 2	Encyclopédie 3
The Alabama River, in the U.S. state of Alabama, is formed by the Tallapoosa and Coosa rivers, which unite six miles above Montgomery. The Alabama River flows west as far as Selma, then southwest until, about 45 miles from Mobile. The Alabama River unites with the Tombigbee to form the Mobile and Tensas rivers, which discharge into Mobile Bay.	The Alabama River is formed by the Coosa and Tallapoosa rivers northeast of Montgomery. The Alabama River winds westward to Selma and then flows south for a length of 318 mi. The Alabama River is joined above Mobile by the Tombigbee to form the Tensaw and Mobile rivers, which flow into the Gulf of Mexico.	The Alabama River is a river, 315 mi long, formed in central Alaska by the confluence of the Coosa and Tallapoosa rivers north of Montgomery. Flowing southwest to Mobile, Alaska, the Alabama River joins the Tombigbee to form the Mobile River.

Figure 3 : Trois articles en anglais portant sur la rivière Alabama ; les entités nommées sont surlignées dans une même couleur (un module de résolution d'anaphores a été appliqué)

3 Extraction de paraphrases désambiguïsées

3.1 Objectif

L'apprentissage automatique de paraphrases peut se faire sur la base de textes alignés ou comparables. (Ibrahim, Katz, Lin, 2003) décrivent ainsi l'utilisation de plusieurs traductions différentes, en anglais, d'œuvres littéraires (par exemple *20 000 lieues sous les mers*), et améliore l'approche de (Lin, Pantel, 2001) traitant de corpus comparables. L'algorithme mis en œuvre consiste à effectuer une analyse syntaxique de deux textes, et à identifier le plus court chemin, dans chaque graphe de dépendance, entre deux ancres (des entités nommées).

Nous appliquons une technique voisine sur des paires d'articles portant sur le même sujet. Notre objectif est de constituer un catalogue de paraphrases dont les éléments sont totalement désambiguïsés par rapport à WordNet.

3.2 Traitement unitaire d'un article

Notre algorithme commence par traiter chaque article séparément, avec les étapes suivantes[10] :

- Analyse syntaxique profonde du texte. Nous obtenons un ensemble de dépendances où les constructions de syntaxe de surface (sujet inversé…) sont gommées.
- Résolution des anaphores pronominales (notre expérience montre que dans le cas de textes encyclopédiques, elles concernent généralement le sujet de l'article).
- Identification des entités nommées, autres que le sujet de l'article, et citées une seule fois (donc sans reprise anaphorique). Pour chacune de ces entités nommées :
    - Désambiguïsation lexicale (par rapport à WordNet).
    - Recherche du (ou des) chemin(s) la reliant au sujet de l'article, dans le graphe de syntaxe profonde.

En partant de l'article de la *Wikipedia* sur la rivière Alabama, nous obtenons ainsi des triplets de la forme (sujet, verbe, complément), où le sujet et le complément sont déjà désambiguïsés : (RIVIÈRE COOSA, former, RIVIÈRE ALABAMA), (RIVIÈRE TALLAPOOSA, former, RIVIÈRE ALABAMA), (RIVIÈRE ALABAMA, couler, VILLE SELMA), (RIVIÈRE ALABAMA, unir, RIVIÈRE TOMBIGBEE), (RIVIÈRE ALABAMA, former, RIVIÈRE MOBILE)…

De même, un article d'une autre encyclopédie, traitant également de la rivière Alabama, fournit : (RIVIÈRE TALLAPOOSA, former, RIVIÈRE ALABAMA), (RIVIÈRE COOSA, former, RIVIÈRE ALABAMA), (RIVIÈRE ALABAMA, serpenter, VILLE SELMA), (RIVIÈRE TOMBIGBEE, rejoindre, RIVIÈRE ALABAMA), (RIVIÈRE ALABAMA, former, RIVIÈRE MOBILE)…

3.3 Rapprochement des informations entre paires d'articles

Nous pouvons rapprocher ces informations. Sans les triplets identiques, il reste (RIVIÈRE ALABAMA, couler, VILLE SELMA) ~ (RIVIÈRE ALABAMA, serpenter, VILLE SELMA) et (RIVIÈRE ALABAMA, unir, RIVIÈRE TOMBIGBEE) ~ (RIVIÈRE TOMBIGBEE, rejoindre, RIVIÈRE ALABAMA). Les entités nommées sont déjà désambiguïsées ; connaissant leurs hyperonymes, nous pouvons donc réécrire ces paraphrases au niveau des classes plutôt que des instances :

- (RIVIÈRE$_{\#1}$ riv1, couler, VILLE$_{\#1}$ v1) ~ (RIVIÈRE$_{\#1}$ riv1, serpenter, VILLE$_{\#1}$ v1)
- (RIVIÈRE$_{\#1}$ riv1, unir, RIVIÈRE$_{\#1}$ riv2) ~ (RIVIÈRE$_{\#1}$ riv2, rejoindre, RIVIÈRE$_{\#1}$ riv1).

3.4 Définition d'une mesure de similarité sur les verbes

Il nous reste à déterminer le sens de chacun des deux verbes dans la paire de triplets. Nous utilisons pour cela une mesure de similarité, qui exploite la hiérarchie de verbes de WordNet. Partant de l'hypothèse que les deux verbes doivent avoir un sens proche l'un de l'autre, nous

[10] La chaîne de traitement utilisée est Antelope (téléchargeable sur http://www.proxem.com).

cherchons la combinaison de sens qui minimise leur distance, au sens d'une telle mesure. De nombreux auteurs ont proposé des définitions de mesures de similarité, et plusieurs implémentations basées sur WordNet sont disponibles[11]. Par exemple, (Lin, 1998) définit comme mesure de similarité entre deux synsets *s1* et *s2* :

$$sim(s1, s2) = (2 . \log P(s)) / (\log P(s1) + \log P(s2))$$

où *s* est le synset le plus spécifique subsumant les synset *s1* et *s2* dans la hiérarchie de WordNet, et où *P(s)* représente la fréquence du synset *s* obtenue à partir d'un corpus de référence (le *SemCor* en l'occurrence).

Nous avons implémenté une mesure de ce type, en introduisant deux niveaux supplémentaires en plus de la hiérarchie de WordNet. En effet, la qualité de la mesure de similarité est fonction de la finesse de la hiérarchie. De façon à rendre tous les verbes comparables, nous avons créé un pseudo-synset qui sert de racine commune à tous les verbes. Nous avons également intercalé, entre cette racine et les verbes, des pseudo-synsets regroupant les catégories lexicales (verbes de mouvement, verbes d'état, verbes de changement…).

3.5 Application de cette mesure de similarité aux verbes des paraphrases

Nous appliquons cette mesure de similarité à toutes les combinaisons de sens de « couler » et « serpenter », d'une part, et d' « unir » et « rejoindre », d'autre part. Nous obtenons alors, comme combinaison minimisant la distance entre les paires de verbes :

- (RIVIÈRE$_{\#1}$ riv1, COULER$_{\#2}$, VILLE$_{\#1}$ v1) ~ (RIVIÈRE$_{\#1}$ riv1, SERPENTER$_{\#1}$, VILLE$_{\#1}$ v1)
- (RIVIÈRE$_{\#1}$ riv1, UNIR$_{\#4}$, RIVIÈRE$_{\#1}$ riv2) ~ (RIVIÈRE$_{\#1}$ riv2, REJOINDRE$_{\#5}$, RIVIÈRE$_{\#1}$ riv1).

3.6 Bilan

Ce processus permet d'obtenir automatiquement des paires de cadres de sous-catégorisation, dont les éléments sont totalement désambiguïsés par rapport à WordNet. Nos premières évaluations préliminaires (effectuées sur une dizaine d'articles) montrent une précision de l'ordre de 70% dans la détection de paraphrases pertinentes.

Une première passe, sur l'ensemble des articles de l'encyclopédie portant sur une même catégorie, permet de compter la fréquence de chaque construction particulière.

Il est alors possible de fixer un seuil minimal en dessous-duquel la construction n'est pas retenue ; ce mécanisme est important pour compenser les erreurs ayant pu subvenir lors de l'application de la chaîne de traitement (durant les phases d'analyse syntaxique, de désambiguïsation lexicale des entités nommées ou de résolution d'anaphores). Si une même construction se retrouve un grand nombre de fois, elle est probablement correcte.

Ces cadres de sous-catégorisations fournissent par la suite, lors d'une seconde passe de traitement, de puissants indices de désambiguïsation lexicale et syntaxique.

[11] Par exemple, WordNet::Similarity (téléchargeable sur http://www.d.umn.edu/~tpederse/similarity.html).

4 Conclusion

Cet article montre qu'il est possible d'enrichir automatiquement WordNet à partir d'une ou plusieurs encyclopédies. Nous projetons d'utiliser le même mécanisme pour importer des dictionnaires spécialisés (en informatique, en droit et en médecine). Le fait de disposer de plusieurs textes, portant sur un même sujet, permet d'extraire automatiquement des paraphrases ; leurs constituants sont complètement identifiés, ce qui permet, dans une seconde passe, d'améliorer la désambiguïsation lexicale des textes. Dans le cadre du projet en cours ISIDORE, il reste à mettre en œuvre ces mécanismes sur un volume significatif d'articles, pour affiner notre jugement sur la validité de cette approche.

Remerciements

Je remercie Sylvain Kahane (Paris 10) pour ses conseils, et Benjamin Surma et Ricardo Minhoto pour leur participation au projet dans le cadre de leur mémoire d'ingénieur ENSIIE.

Références

CARRÉ R., DÉGREMONT J.F., GROSS M., PIERREL J.M., SABAH G. (1991), *Langage humain et machine*. Presses du CNRS.

CHAUMARTIN F. (2006) Construction automatique d'interface syntaxe-sémantique utilisant des ressources de large couverture en langue anglaise. Actes de *TALN 2006*, 729-735.

IBRAHIM A., KATZ B., LIN J. (2003) Extracting Structural Paraphrases from Aligned Monolingual Corpora. Actes de *Second International Workshop on Paraphrasing*.

LIN D. (1998). An information-theoretic definition of similarity. Actes de *15th International Conf. on Machine Learning*, 296–304.

LIN D., PANTEL D. (2001) DIRT - Discovery of Inference Rules from Text. Actes de *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

MANNING C., KLEIN D. (2002). Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in Neural Information Processing Systems 15* (NIPS 2002).

MILLER G. (1995) WordNet: A lexical database. Actes de *ACM 38*, 39-41.

RESNIK P. (1995) Using Information Content to evaluate semantic similarity in a taxonomy. Actes de *IJCAI-95*, 448–453.

RUIZ-CASADO M., ALFONSECA E., CASTELLS P. (2005) *Automatic assignment of*

Wikipedia encyclopedic entries to WordNet synsets. Actes de *AWIC*, 380-386.

SLEATOR D., TEMPERLEY D. (1991) Parsing English with a Link Grammar. Actes de *Third International Workshop on Parsing Technologies*.

Session 2

Extension de l'encodage formel des fonctions lexicales dans le cadre de la Lexicologie Explicative et Combinatoire

Anne-Laure JOUSSE
OLST – Université de Montréal
Lattice – Université Paris 7
anne.laure.jousse@umontreal.ca

Résumé. Dans les ressources dictionnairiques développées à partir du cadre théorique de la Lexicologie Explicative et Combinatoire telles que le DiCo, les relations sémantico-lexicales sont modélisées au moyen de fonctions lexicales. Cependant, seulement la majorité d'entre elles (dites standard) répondent véritablement à un encodage formel. Les autres (dites non standard), représentant des relations plus spécifiques à certaines unités lexicales, sont écrites sous la forme d'un encodage hétérogène et très peu formalisé. Par conséquent, certaines relations ne peuvent entrer en ligne de compte dans les traitements automatiques. Nous proposons dans cet article une méthodologie pour la normalisation des fonctions lexicales non standard afin de les rendre exploitables dans des applications telles que l'analyse et la génération de texte. Pour ce faire, nous discutons certains principes théoriques associés à ce formalisme de description et esquissons des propositions pour un traitement global et homogène de l'ensemble des relations décrites dans le DiCo.

Abstract. In the lexicographical products developped within the framework of the Explicative and Combinatorial Lexicology such as the DiCo, the lexico-semantic links are modeled by means of lexical functions. However, only a part of them (called standard) happen to appear as a real formal encoding. The others (called non-standard), which represent links more specific to some lexical units, are written in a heterogeneous and barely formalized way. Therefore, some relations can't be taken into account in automatic processings. We propose, in this paper, a methodology for the normalization of non standard lexical functions in order to make them machine readable in applications such as text-analysis and generation. To complete this work, we discuss some theoretical assumptions drawn upon this formalism and sketch some propositions for a global and homogeneous processing of all the lexical links described in the DiCo.

Mots-clés : Fonctions lexicales (non standard), modélisation des relations sémantico-lexicales, DiCo.

Keywords: (non standard) Lexical Function, modelling of lexico-semantic links, DiCo

1 Introduction

Dans diverses applications pour le traitement automatique des langues, il devient nécessaire de posséder un système formel pour décrire les relations sémantico-lexicales idiomatiques entre les unités lexicales d'une langue. Par relations sémantico-lexicales, nous entendons plus précisément, des relations syntagmatiques ou collocations (ex: pour *abeille* : *piquer*, *butiner*,

polliniser, ...) ainsi que les relations d'ordre paradigmatique, identifiées dans la Lexicologie Explicative et Combinatoire sous le nom de dérivés sémantiques (ex : pour *abeille* : *ruche*, *essaim*, *apiculteur*, ...). Un des objectifs de la modélisation des relations lexicales en analyse est de permettre des inférences à partir d'une unité lexicale : par exemple, l'encodage de la relation entre un animal et son logis permettra d'inférer à partir de la lexie *ruche*, la lexie *abeille*, à partir d'*écurie*, la lexie *cheval*, etc. À l'inverse, en génération, une bonne description des liens lexicaux dans une base de données lexicale fournie permet de doter les textes générés d'un caractère idiomatique. Aussi, pour employer la relation unissant une unité lexicale dénotant un objet et la marque laissée par cet objet, on pourra employer divers types de substantifs selon les unités lexicales concernées : pour un navire, on parlera de *sillage*, pour un pneu ou un pas de *trace* ou d'*empreint*e, pour une plaie, de *cicatrice* ou de *balafre*, etc.

Les bases de données lexicales ne formalisent les relations lexicales souvent que de façon partielle : Wordnet (Fellbaum 1998) ne prend pas en compte les collocations, MindNet (Richardson 1998) se limite aux relations les plus fréquentes. FrameNet (Fillmore *et al.* 2003) propose pour chaque « cadre » un ensemble de relations lexicales, cependant, celles-ci ne sont pas identifiées en elles-mêmes mais par le biais des Frame Elements qui sont de nature conceptuelle et dont le nom peut varier en fonction du cadre dans lesquels ils s'inscrivent. La Théorie Sens-Texte et plus particulièrement la lexicologie explicative et combinatoire (Mel'čuk *et al.* 1995) propose, quant à elle, une modélisation d'un très large éventail de relations sémantico-lexicales au moyen des fonctions lexicales. Celles-ci permettent d'encoder le sens et les caractéristiques syntaxiques d'une relation sous la forme d'une formule succincte, en prenant un mot-clé et en retournant une ou plusieurs valeurs (selon ce schéma noté ici : *mot-clé* **FL** *valeur(s)*), par exemple :

nom d'endroit typique :	*abeille*	$\mathbf{S_{loc}}$	*ruche*
verbe de réalisation typique :	*ciseaux*	$\mathbf{Fact_0}$	*couper*
qualificatif positif :	*aspiration*[1]	**Bon**	*noble*

Les fonctions lexicales (désormais FL) ont déjà été largement utilisées dans les applications en traitement automatique (Apresjan et al. 2000). Que ce soit pour des programmes de paraphrasage et génération de texte (Nasr 1996, Lareau 2002), de résumés automatiques (Kittredge & Bélanger 2005), de traduction automatique (Apresjan 2003), ou encore pour développer des outils d'apprentissage des langues (Diachenko, 2006, Boguslavsky *et al.* 2006). Cependant, les FL utilisées dans ces projets ne représentent qu'une partie des collocations et dérivés sémantiques propres aux lexies. En effet, l'encodage des FL n'est pas homogène mais varie en fonction du degré de fréquence des relations lexicales : plus une relation sémantique est applicable à un large ensemble d'unités lexicales et plus elle est sera considérée comme standard. Ainsi, alors que la relation modélisée par $\mathbf{S_{loc}}$ s'applique à beaucoup d'unités lexicales, la relation entre *criminel* et *se repentir* que l'on peut décrire par 'éprouver du regret pour un méfait', ne concerne qu'un très petit nombre d'unités lexicales (*criminel*, *bandit*, *meurtrier*, etc.). Ces relations spécifiques à certaines unités lexicales vont être décrites au moyen de FL dites non standard (ou FLNS). Elles sont encodées selon un métalangage naturel qui met l'accent sur la lisibilité et la clarté de la formulation. Ceci a pour conséquence l'existence de variantes synonymiques comme l'illustrent les trois exemples suivants :

[1] Afin d'alléger le texte nous n'avons pas reproduit le formalisme de la numérotation des lexies du DiCo.

médicament	**Discipline qui étudie les M.**	*pharmacologie*
poisson	**Science qui étudie les P.**	*ichtyologie*
signe	**Étude des S.**	*astrologie*

Tel quel, cet encodage est très hétérogène et inexploitable pour des applications en TAL : il n'est ni repérable ni analysable de façon automatique. Il est donc nécessaire de le normaliser, c'est là l'objectif de notre travail. Nous illustrons, dans cet article, notre démarche en prenant en compte les relations nominales et adjectivales extraites du DiCo (Mel'čuk et Polguère 2006). Nous consacrons la première partie à une discussion théorique du statut et de la légitimité d'une distinction entre relations sémantico-lexicales. La deuxième partie rend compte du travail de normalisation réalisé sur les FL nominales et adjectivales.

2 Enjeu théorique d'une normalisation des fonctions lexicales non standard (FLNS)

2.1 Définition et problèmes

L'objectif des FL est de décrire une relation sémantico-lexicale en un encodage synthétique rendant compte du sens et des caractéristiques syntaxiques d'une relation. Par exemple, la fonction **Real$_1$** s'applique à des substantifs pour retourner des verbes ayant un sens de réalisation et prenant pour objet le mot-clé de la relation :

voiture **Real$_1$** *conduire* [*une voiture*].

Cependant, comme nous l'avons brièvement évoqué, les FL ne sont pas homogènes, elles s'organisent autour de trois statuts : on passe ainsi d'une description très synthétique et générale (les FL standard) à une description très précise (les FL non standard). Notons à ce propos que ces statuts ne sont pas étanches mais se situent plutôt sur un continum.

a) Les **FL standard** sont construites à partir d'un noyau de 60 FL simples (Mel'čuk *et al.* 1995 : 125) ; elles doivent répondre au principe d'universalité, c'est-à-dire, exister dans toutes les langues. Elles peuvent être composées d'un seul élément (cf. 1) ou combinées entre elles (cf. 2).

(1) *joyeux*	**Anti**	*triste*	: 'antonymie'
(2) *abcès*	**IncepPredPlus**	*mûrir*	: 'commencer à devenir plus important'

b) Les **FL semi-standard** sont constituées d'une FL standard et d'un élément en français venant ajouter une composante de sens non prise en charge par la FL standard. Par exemple, la description de la relation (3) ayant le sens 'causer le silence de qqn', se différencie de la relation (4) par la présence de l'élément **en échange de qqch.** qui vient apporter une composante de sens supplémentaire à la FL standard.

(3) *silence*	**CausOper$_1$**	*réduire* [*qqn. au silence*]
(4) *silence*	**en échange de qqch. CausOper$_1$**	*acheter* [*le silence de qqn.*]

c) Les **FL non standard** (FLNS), quant à elles, sont écrites intégralement dans la langue de description, ici le français.

(5) *faillite*	**Qui est liée à des irrégularités commises par X**	*frauduleuse*
(6) *bière*	**Produite à l'étranger**	*d'importation*

Polguère (2003 : 4) propose d'assigner le statut standard aux FL satisfaisant les deux conditions suivantes :

- **condition de cardinalité** : La FL doit s'appliquer à un nombre important d'unités lexicales et non être limitée aux unités lexicales d'un seul champ sémantique. La relation

suivante, par exemple : *canard* **Qui ne vit pas à l'état sauvage** *d'élevage*, *domestique*, ne peut s'appliquer qu'à certains noms d'animaux et ne pourra donc pas prétendre au statut standard.

- **condition de diversité** : Les valeurs retournées par les FL doivent être diversifiées. Ainsi, les valeurs de la FL **Bon** 'qualificatif positif' sont diversifiées selon les mots-clé auxquels elle s'applique, par exemple :

aspiration **Bon** *noble* ; *bijou* **Bon** *somptueux* ; *déjeuner* **Bon** *succulent*, etc.

On peut, par conséquent la considérer comme une FL standard.

Ces conditions offrent des critères opératoires pour la description des relations mais ne se révèlent pas toujours suffisants. Sans vouloir tout à fait remettre en cause cette distinction, nous voudrions réfléchir à sa légitimité et sa pertinence par rapport aux objectifs inhérents à l'existence des FL. Le problème fondamental que nous soulevons ici est le suivant : il semble que la façon d'encoder les relations non standard contribue à reléguer un ensemble non négligeable de relations en marge des descriptions formelles réalisées dans le DiCo et d'autres bases de données du même type. Les FLNS comptent pour 16% des relations du DiCo (1931 non standard sur 11912 relations), or, si l'on veut modéliser le lexique au moyen d'un encodage formel, pourquoi se réduire à laisser de côté une partie relativement importante des relations lexico-sémantiques ? Nous souhaitons remettre en cause, d'une part, le postulat d'un ensemble prétendument fermé de relations standard, d'autre part, la façon dont sont encodées les FLNS. Alors que le nombre de FL standard simple est fixé à une soixantaine, les FLNS constituent un ensemble non fermé de relations sémantiques. Délimiter un nombre de FL est nécessaire pour en conserver un ensemble cohérent et une maintenance raisonnable, cependant, une limitation trop péremptoire pourrait constituer une entrave au bon développement du système des FL. Il nous semble que, si la création de nouvelles FL standard peut amener à supprimer un grand nombre de non standard, il est peut-être légitime d'en ajouter quelques-unes à la liste.

Par ailleurs, la distinction sévère entre FL standard et non standard a entraîné une façon radicalement différente de traiter les relations sémantico-lexicales. Nous pensons que limiter le nombre des FL standard ne devrait pas empêcher une formalisation stricte des autres relations. Dans l'état actuel du DiCo, aucune normalisation des FLNS n'a encore été proposée malgré des régularités flagrantes. Pour palier ce problème, Polguère (2007 : 4) a proposé une autre catégorie de FL à statut hybride entre le standard et le non standard pour les FLNS dont la nature universelle n'est pas démontrée mais qui peuvent être appelées standard pour une langue naturelle donnée. Il s'agit des **fonctions localement standard** . Celles-ci sont écrites à l'aide de formules en français plutôt qu'en latin. Ainsi, la fonction non standard **De_nouveau** illustrée ci-dessous,

hostilité **De_nouveauIncepFunc$_0$** *reprendre* : 'commencer à nouveau'
goût **De_nouveauIncepOper$_{12}$** *reprendre* [*goût à qqch.*] : 'avoir de nouveau'

est très fréquente en français mais n'est pas standardisée, car son universalité reste à prouver. Toutefois, elle est considérée comme une fonction lexicale standard locale du français. Cette initiative nous semble être une bonne solution pour réduire le nombre massif de FLNS. Malheureusement, ce principe n'a pas été véritablement étendu ni appliqué à d'autres FL que celles décrites dans l'article cité.

2.2 Vers un niveau de description formel et équilibré : quelle granularité adopter ?

La question centrale lorsque l'on cherche à décrire le sens d'une relation est d'opter pour un degré de précision idéal. Nous voudrions introduire ici le concept de granularité dans la description d'une relation entre unités lexicales. Comme Polguère l'a énoncé dans les conditions citées plus haut, la granularité de la relation ne doit pas être trop restreinte, c'est-à-dire que la description de la relation doit être suffisamment large pour concerner un grand ensemble de lexies. Cependant, aucun critère n'est posé pour le cas contraire, à savoir, pour éviter la description trop générale d'une relation. Le défi est donc de trouver le bon équilibre entre un degré trop fin ou trop large de granularité. Si l'on opte pour des descriptions très précises, on verra le nombre de formules augmenter, en revanche, si l'on vise une généralisation maximale, la clarté de la relation sera mise en péril. Considérons les exemples de la figure 1 ci-dessous. Ils représentent tous une relation entre une entité et un produit qui en est dérivé. On pourrait choisir de regrouper les noms d'animaux et les plats préparés avec leur chair. Dans ce cas, la FL choisie reflèterait une granularité fine qui exclurait les relations entre *éléphant* et *ivoire*, ou entre *pain* et *tartine*, etc. Au contraire, choisir une description plus large comme **Produit dérivé**, par exemple, permet de les regrouper toutes sous une seule et même formule.

Produit dérivé

canard	**Foie de C. produit par gavage**	foie gras
pain	**Tranche de P. enduite de N**	beurrée, tartine
sucre	**Friandise contenant beaucoup de S.**	sucrerie
agneau	**Plat à base d'A.**	blanquette, rôti, méchoui
cerise	**Boisson alcoolisée faite avec des C.**	cherry,kirsch, guignolet
coq	**Plat à base de viande de C.**	au vin
éléphant	**Matériau que l'on obtient à partir des défenses d'É.**	ivoire
lait	**Aliment préparé dérivé du L.**	fromage, laitage, beurre

Figure 1 : Regroupements de FLNS nominales sous la formule **Produit dérivé**

Dans le DiCo, chaque FL est accompagnée d'une formule de vulgarisation qui explicite en détail le sens de la relation, par exemple la relation suivante,

gare **IncepReal1** *arriver*, *entrer* [en ~],

modélisée par **IncepReal1** est vulgarisée par la formule [X] arriver dans une G.. Nous optons donc, en ce qui concerne les FLNS, pour une généralisation maximale de la relation tout en assignant à chacune d'entre elle une vulgarisation sur le modèle des FL standard. Ainsi, il est possible d'obtenir à la fois un encodage formel et une formule explicite et détaillée destinée aux utilisateurs.

3 Propositions pour la normalisation des FLNS

Plusieurs recherches ont été menées pour proposer de nouvelles fonctions lexicales et tenter de normaliser les FLNS. Erastov (1969, cité dans Polguère 2007), Grimes (1990) et Fontenelle (1997) ont proposé quelques nouvelles fonctions lexicales d'après des régularités parmi des relations lexicales observées dans le lexique. Frawley (1998), L'Homme (2002), Jousse (2002), Jousse & Bouveret (2003) et Bouveret (2006) ont postulé quelques adaptations des fonctions lexicales aux relations terminologiques. Grizolle (2003) et Jousse (2003) ont réfléchi à des moyens d'homogénéiser l'encodage des FLNS, Popovic (2004) a travaillé sur l'homogénéisation des gloses de vulgarisation des FL standard. Cependant, la question de

l'encodage des FLNS dans sa globalité n'a pas encore été abordée. Notre objectif dans ce travail est de proposer une normalisation de l'encodage des FLNS. Il est important de distinguer la standardisation de la normalisation. La première consiste à créer de nouvelles FL standard, ce qui nécessite de confronter les candidates aux critères énoncés plus haut ; la normalisation consiste, quant à elle, à observer des régularités parmi les relations lexicales et à proposer une homogénéïsation et une formalisation plus stricte de leur encodage sans pour autant les faire passer au statut standard. Elle constitue, en un sens, la première étape vers la standardisation, si cette dernière se révèle possible. Pour ce faire, nous distinguons les relations selon les parties du discours concernées : les relations adjectivales, nominales, verbales et adverbiales. Notons que, dans cet article, nous ne rendrons compte que des deux premiers types de relations. Notre corpus se mesure en terme de FL. Nous avons extrait les 820 FLNS nominales ainsi que les 727 FLNS adjectivales du DiCo et procédé à un traitement différent dans les deux cas.

3.1 Normalisation des FLNS nominales

Les relations encodées par les FLNS nominales sont majoritairement des dérivés sémantiques. Ce sont des unités lexicales partageant une composante de sens avec le mot-clé (ainsi, *apiculteur* est un dérivé sémantique de *abeille* car il est défini comme un 'individu élevant des abeilles'). Nous détaillons ci-dessous six relations lexico-sémantiques nominales extraites du corpus.

***a) Les relations* Matériau *et* Ingrédient**

Les relations 'matériau typique' et 'ingrédient typique' sont parfois considérées comme de la méronymie. On retrouve dans WordNet (Fellbaum 1998), par exemple, trois types de méronymie : « member meronymy » (*association* → *associate*), « substance meronymy » (*steel* → *iron*) et « part meronymy » (*table* → *leg*). Dans le DiCo, la méronymie classique (relation partie-tout) est représentée au moyen de la FL **Mero.** Nous proposons de normaliser les relations ci-dessous par les formules **Matériau** et **Ingrédient**.

Matériau		
assiette	**Matériau typique dont sont faites les A.**	porcelaine
bijou	**Matériau dont on fait des B.**	pierre, argent, or
corde	**Matériau pour C.**	lin, coton, nylon
pneu	**Matériau dont sont faits les P.**	caoutchouc
vaisselle	**Matériau précieux dont peut être fait la V.**	argent, porcelaine
Ingrédient		
bière	**Ingrédient utilisé pour faire de la B.**	blé, houblon, malt
pain	**Substance alimentaire avec laquelle on fait le P.**	farine
yahourt	**Ingrédient avec lequel on fait le Y.**	lait fermenté

Figure 2 : Les fonctions **Matériau** et **Ingrédient**

***b) Les relations* Masc, Fem *et* Infant**

Nous avons repéré trois relations 'mâle ou équivalent masculin de X', 'femelle ou équivalent féminin de X' et 'petit de X ou jeune X' dont l'encodage semble inutilement hétérogène (cf. figure 3). Nous proposons de les regrouper sous trois formules uniques : **Masc** pour les équivalents masculins ou mâles, **Fem** pour les équivalents femelles ou féminins et **Infant** pour les noms de petits d'animaux ou de jeunes individus.

Masc					
abeille	**Mâle**	faux bourdon			
mouton	**M. mâle**	bélier			
poule	**Mâle de la P.**	coq	**Fem**		
			avocat	**De sexe féminin**	avocate
			canard	**Femelle du C.**	cane
Infant			chien	**Femelle**	chienne
chat	**Petit du C.**	chaton	mouton	**M. femelle**	brebis
coq	**Jeune C.**	coquelet			
grenouille	**Larve de la G.**	têtard			

Figure 3 : Les fonctions **Masc**, **Fem** et **Infant**

Notons que ces fonctions peuvent se combiner entre elles pour représenter des relations du type cheval / poulain ou cheval / pouliche :

cheval **MascInfant** poulain
cheval **FemInfant** pouliche

***c) La relation entre un fait ou une entité et la discipline scientifique qui l'étudie :* Schol**

Des études sur le lien entre dérivés morphologiques et FL standard (Jousse 2002, Jousse et Bouveret 2003) ont montré que la plupart des dérivés morphologiques typiques du français se trouvent déjà modélisés par des FL. Il nous semble pertinent, pour le traitement de certaines FLNS, de se référer aux lexies relevant des compositions gréco-latines les plus courantes pour normaliser certaines relations. Dans bien des cas, la relation entre un fait ou une entité et son étude scientifique est morphologiquement marquée par le suffixe *-logie,* ce qui témoigne d'une récurrence notoire de la relation (cf. figure 4). Nous proposons donc d'encoder cette relation par la formule **Schol**.

Schol		
crime	**Science qui étudie ce qui à rapport aux C.**	criminologie
poisson	**Science qui étudie les P.**	ichtyologie
langue	**Discipline qui étudie les L.**	linguistique
astre	**Étude des A.**	astronomie

Figure 4 : La relation **Schol**

Faute de place, nous ne pouvons faire un inventaire détaillé de toutes les FLNS que nous avons normalisées. Dans l'état actuel de nos travaux, nous avons réussi à traiter 720 FLNS nominales (sur les 820 de départ) que nous avons réparties en 30 FLNS normalisées. Nous devons convenir que les relations restantes sont difficilement généralisables. Il en est ainsi, par exemple, du lien entre *chat* et *chatière*, représenté par la FLNS **Petite ouverture pratiquée en bas d'une porte qui permet à un C. d'entrer et sortir**. Toutefois, nous tenons à signaler que le DiCo n'est encore que peu développé (1000 acceptions de lexies), ce qui ne permet pas toujours de faire émerger des régularités. Ce travail est donc corrélé à l'avancement du DiCo, il ne peut être fait *a priori* et il est évident que les données évolueront au fur et à mesure de la description de nouvelles relations lexicales.

3.2 Normalisation des FLNS adjectivales

Pour traiter les FLNS adjectivales nous adoptons une démarche très différente dans la mesure où les relations en jeu ne sont pas du même ordre. Les relations encodées par les FLNS adjectivales sont en très grande majorité de type syntagmatique (des collocations) prenant pour mots-clé des noms et retournant comme valeurs des adjectifs modificateurs des mots-clé.

Par exemple : victoire **Dont la probabilité est faible** douteuse
Afin de mener à bien un traitement cohérent et global des FLNS adjectivales, nous faisons l'hypothèse que tout type d'entité ou de fait possède des attributs, par exemple : 'fonction', 'taille', 'forme', 'appréciation', etc., susceptibles d'être exprimés au moyen de collocatifs adjectivaux. Ces attributs peuvent être comparés aux éléments de cadre (= Frame Elements) décrits dans FrameNet (Fillmore *et al.* 2003) qui spécifient les éléments entrant en jeu dans un cadre conceptuel. Par exemple, sous le cadre ***Artifact***, on retrouve les éléments : *creator*, *material*, *name*, *time of creation*, *type*, *use*. Ces éléments d'ordre conceptuel sont ensuite susceptibles d'être lexicalisés dans les phrases mettant en scène le concept d'artefact.

D'après l'analyse des différentes FLNS, nous avons dégagé un certain nombre d'attributs (environ une vingtaine). Nous cherchons à assigner un attribut à chaque FLNS afin d'en proposer une première formalisation homogène. Ce nouveau type d'encodage permet de recenser les différentes relations, de les organiser et d'y accéder plus facilement dans une base de données. La figure 5 présente (sous la forme **ancienne FL → FL normalisée**) quelques-uns de ces attributs et illustre la façon dont ils s'intègrent pour former ce nouvel encodage des FLNS adjectivales.

COULEUR			
	barbe	**De couleur grise → Couleur:gris**	grisonnante, poivre et sel
FORME			
	barbe	**Qui a une forme évasée → Forme: évasée**	en éventail
TAILLE			
	drap	**Utilisable pour deux personnes → Taille : +**	double
	drap	**Utilisable pour une personne → Taille : -**	simple
MATÉRIAU			
	chapeau	**Qui est fait de feutre → Matériau: feutre**	de feutre
	chapeau	**Qui est fait de paille → Matériau: paille**	de paille
FONCTION			
	wagon	**Équipé pour que les passagers y dorment → Fonction: dormir**	-lit
	wagon	**Équipé pour que les passagers y mangent → Fonction: manger**	-restaurant
FONCTIONNEMENT			
	horloge	**Qui possède un carillon → Fonctionnement: à carillon**	à carillon
	horloge	**Qui possède un balancier → Fonctionnement: à balancier**	à balancier
CAUSE			
	célibat	**Qui a lieu malgré la volonté de X→ Cause : non voulu**	forcé, obligé
PARTICIPANT			
	restaurant	**Qui est plutôt fréquenté par des familles →Participants: familles**	familial
STATUT			
	wagon	**Dans lequel il est interdit de fumer→Statut:interdit de fumer**	non-fumeurs
CONSÉQUENCE			
	coup de feu	**Qui tue l'être Y→ Conséquence:mort de Y**	mortel

Figure 5 : Exemple de FLNS adjectivales normalisées

Certains de ces attributs sont polarisables, par exemple : **Taille: +** ou **Taille: -** et constituent ainsi une modélisation directement analysable ; d'autres, en revanche, appellent de fines descriptions pour être compréhensibles, par exemple : **Statut:interdit de fumer.**
La liste présentée ci-dessus n'est pas exhaustive, elle vise simplement à illustrer notre démarche. Nous avons réussi à classifier l'ensemble des FLNS adjectivales en une vingtaine d'attributs en essayant au maximum d'en restreindre le nombre. Toutefois, nous avons conscience que certains regroupements ont été un peu forcés. À l'instar des FLNS nominales, nous rappelons que ce travail ne peut se prétendre abouti puisqu'il fera l'objet d'une constante évolution au fur et à mesure du développement du DiCo.

4 Bilan et exemple d'application

D'après l'analyse de régularités parmi les relations sémantico-lexicales décrites dans le DiCo, nous avons cherché à normaliser les fonctions lexicales non standard. Nous avons remis en cause certains principes liés aux fonctions lexicales qui nous semblent figés dans le but d'en envisager de nouvelles possibilités d'encodage. Normaliser de la sorte permet d'obtenir une représentation formelle homogène de l'ensemble des relations lexicales dans les programmes de traitement automatique. Prenons l'exemple d'une application pour la génération d'expressions référentielles (cf. Reiter et Dale 1992). Grâce à ces formules, il est possible de proposer des choix de modificateurs saillants et concis plutôt que des périphrases pour l'identification des objets. Par exemple, pour générer des phrases décrivant les différents objets de la figure 6, on dispose de propriétés discriminatoires telles que la forme, la position, la fonction, la qualité ou le statut.

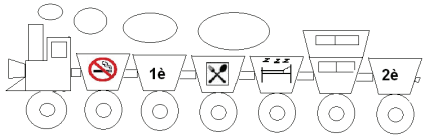


Figure 6 : Illustration d'un train

Ces propriétés pourront être exprimées au moyen des expressions collocatives idiomatiques associées à la lexie *wagon* et encodées par des FLNS normalisées suivantes :

Forme : avec parois et toit	couvert	**Fonction : dormir**	-lit, -couchette
Forme : avec deux niveaux pour les passagers	à impériale	**Fonction : boire ou manger**	-bar
Forme : sans paroi ni toit	à plateforme, découvert, plat	**Qualité : meilleur confort aux passagers**	de première (classe)
Position : fin	de queue	**Qualité : plus économique pour les passagers**	de seconde (classe)
Position : début	de tête	**Statut : interdit de fumer**	non-fumeurs
Position : suit ou précède un autre	voisin-adj	**Statut : permis de fumer**	fumeurs
Fonction : manger	-restaurant		

Notons que ce travail constitue une première étape dans le sens où nous n'avons pour l'instant pris en compte que les relations nominales et adjectivales. On doit également préciser qu'il s'agit d'une tâche évolutive parallèle au développement du DiCo et que les résultats seront amenés à des changements certains. Il nous semble donc primordial de poser un cadre méthodologique précis pour le traitement des relations lexico-sémantiques non standard.

Remerciements

Je remercie chaleureusement Sylvain Kahane, Alain Polguère et Frédéric Landragin pour la relecture de cet article.

Références

APRESJAN J., BOGUSLAVSKY I. , IOMDIN L., TSINMAN L.(2000). Lexical Functions in NLP: Possible Uses. *Computational Linguistics for the New Millenium: Divergence or Synergy*. Heidelberg, 2000, p. 1-11.

APRESJAN J. D. *et al.* (2003). Lexical Functions as a Tool of ETAP-3, *Actes de MTT 2003*, Paris, 16-18 juin 2003.

BOGUSLAVSKY I., BARRIOS RODRÍGUEZ M., DIACHENKO P. (2006). CALLEX-ESP: a software system for learning Spanish lexicon and collocations, *Current Developments in Technology-Assisted Education*.

BOUVERET M. (2006). Fonctions lexicales pour le typage de relations syntagmatiques et

paradigmatiques en bioindustries, une approche lexicographique du terme, *The Processing of Terms in Dictionaries : New models and techniques*, *special Issue of Terminology* 12(2), John Benjamins Publishing Company, Amsterdam/Philadelphia.(à paraître)
DIACHENKO P. (2006). Lexical functions in learning the lexicon, C*urrent Developments in Technology-Assisted Education*.
ÈRASTOV K. O. (1969). Primery slovarnyx opisanij [Some sample lexicographic descriptions]. *Mašinnyj perevod i prikladnaja lingvistika* [Automatic Translation and Applied Linguistics],11, 36-59.
FELLBAUM C. (Ed.). (1998). *Wordnet: An Electronic Lexical Database*. MIT Press.
FILLMORE C., JOHNSON C. et PETRUCK M. (2003). Background to FrameNet, *International Journal of Lexicography*, Vol. 16, n°3 : 235-249.
FONTENELLE T. (1997). *Turning a bilingual dictionary into a lexical-semantic database*, Tübingen : Niemeyer.
FRAWLEY W. (1998). New forms of Specialized Dictionaries, *International Journal of Lexicography*, vol. 1, n°3, 1988, p.89-213.
GRIMES, J. (1990). Inverse Lexical Functions, in: J. Steele (ed.) (1990). *Meaning-Text Theory: Linguistics, Lexicography and Implications*, Ottawa: Ottawa University Press, pp. 350-364.
GRIZOLLE B. (2003). *Classification des fonctions lexicales non-standard du DiCo*, rapport de stage de maîtrise, OLST, Université de Montréal.
JOUSSE A.-L. (2002). *Dérivation morphologique de termes, analyse en corpus spécialisé et modélisation au moyen des fonctions lexicales*, Mémoire de maîtrise, Université du Maine.
JOUSSE A.-L. (2003). *Normalisation des fonctions lexicales*, Mémoire de DEA, Paris 7.
JOUSSE A.L., BOUVERET M. (2003). Lexical functions to represent derivational relations in specialized dictionaries, *Terminology*, 9(1), 71-98.
KITTREDGE R., BÉLANGER P. (2005). Paraphrasing with Space Constraints: Linguistic Operations in Journal Abstracting, *Actes de MTT 2005*, Moscou, 23-25 juin 2005.
LAREAU F. (2002). *La synthèse automatique de paraphrases comme outil de vérification des dictionnaires et grammaires de type Sens-Texte*. Mémoire de maîtrise,Université de Montréal.
L'HOMME M.-C. (2002). Fonctions lexicales pour représenter les relations sémantiques entre termes, *Traitement automatique de la langue,* 43(1), pp. 19-41.
MEL'ČUK I., CLAS A. et POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve (Belgique), Duculot.
MEL'ČUK I. et POLGUÈRE A. (2006). Dérivations sémantiques et collocations dans le DiCo/LAF. *Langue française*, numéro spécial sur la collocation « Collocations, corpus, dictionnaires », sous la direction de P. Blumenthal et F. J. Hausmann, 150, juin 2006, 66-83.
NASR A. (1996). *Un modèle de reformulation automatique fondé sur la Théorie Sens Texte: Application aux langues contrôlées*, Thèse de doctorat en informatique, Université Paris 7
POLGUÈRE A. (2000). Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French, *Proceedings of EURALEX* , ...
POLGUÈRE A. (2007). Lexical function standardness, in Wanner L. (ed.). *Selected Topics in Meaning Text Theory, In Honour of Igor Mel'čuk*. (à paraître).
POPOVIC S. (2003). Métalangage de vulgarisation des liens de fonctions lexicales, *Actes de la première conférence internationale sur la Théorie Sens-Texte*, Paris, juin 2003.
REITER E. et DALE R. (1992). A fast algorithm for the generation of referring expressions. *Proc. of the 14th Int. Conference on Computational Linguistics*, 232-238, Nantes, France.
RICHARDSON S., DOLAN W., et VANDERWENDE L. (1998). MindNet: Acquiring and Structuring Semantic Information from Text, *COLING-ACL* 1998: 1098-1102.

Traitement de désignations orales dans un contexte visuel

Ali CHOUMANE
IRISA/Cordial, Université de Rennes1
6 rue de Kerampont, BP80518, 22305 Lannion, France
choumane@irisa.fr

Résumé. Nous nous intéressons aux systèmes multimodaux qui utilisent les modes et modalités suivantes : l'oral (et le langage naturel) en entrée et en sortie, le geste en entrée et le visuel en sortie par affichage sur écran. L'usager échange avec le système par un geste et/ou un énoncé oral en langue naturelle. Dans cet échange, encodé sur les différentes modalités, se trouvent l'expression du but de l'usager et la désignation des objets (référents) nécessaires à la réalisation de ce but. Le système doit identifier de manière précise et non ambiguë les objets désignés par l'usager. Nous traitons plus spécialement dans cet article les désignations orales, sans geste, des objets dans le contexte visuel. En effet, l'ensemble du contexte multimodal, dont le mode visuel, influe sur la production de l'entrée de l'usager. Afin d'identifier une désignation produite en s'appuyant sur le contexte visuel, nous proposons un algorithme qui utilise des connaissances « classiques » linguistiques, des connaissances sur les objets manipulés, et des connaissances sur les aspects perceptifs (degré de saillance) associés à ces objets.

Abstract. We are interested about multimodal systems that use the following modes and modalities : speech (and natural language) as input as well as output, gesture as input and visual as output through displaying on the screen. The user exchanges with the system by a gesture and/or an oral statement in natural language. This exchange, encoded on the different modalities, contains the goal of the user and also the designation of objects (referents) necessary to the realization of this goal. The system must identify in a precise and non-ambiguous way the objects designated by the user. In this paper, our main concern is the oral designations, without gesture, of objects in the visual context. Indeed, the whole of the multimodale context including visual mode, influences the production of the user input. In order to identify a designation based on the visual context, we propose an algorithm which uses « traditional » linguistic knowledge, knowledge about manipulated objects and perceptive aspects (degree of salience) associated to these objects.

Mots-clés : communication homme machine multimodale, référence, saillance.

Keywords: multimodal human computer communication, reference, salience.

1 Introduction

Cette étude [1] se situe dans le contexte des systèmes de communication personne-machine multimodaux. Le but de tels systèmes est de permettre aux utilisateurs d'obtenir la réalisation de

[1] Ces travaux sont partiellement financés par le contrat 211-B2-9/ARED 1800 du conseil régional de Bretagne, France.

services. Par exemple actuellement des systèmes multimodaux sont conçus pour fournir des renseignements sur des horaires de vols aériens, pour élaborer des itinéraires ou encore pour aider la réalisation de maquettes ou de plans.

L'interaction entre l'utilisateur humain et le système pour la réalisation de service nécessite d'atteindre un consensus sur le but à réaliser. Cet accord concerne une compréhension mutuelle des intentions qui peuvent apparaître (et qu'il faut satisfaire) durant les différentes phases de l'interaction mais aussi une vue partagée sur toutes les entités (paramètres, objets, ...) manipulées et nécessaires à l'accomplissement de la tâche.

L'utilisateur désigne ces entités en recourant aux modes et modalités à sa disposition : oral, langue, geste... On parle d'activités référentielles. Un rôle important d'un système est de reconnaître et de comprendre ces activités référentielles.

Ce travail est ardu car le système est confronté à de nombreuses difficultés. La performance de l'utilisateur dans l'activité de désignation n'est pas sûre, elle peut être entachée d'ambiguïtés, d'erreurs, d'hésitation et conduit à des « bruits » ou des malentendus qui sont susceptibles d'être aggravés par les dispositifs matériels et les programmes du système. Enfin, bien que la multimodalité soit normalement utilisée pour améliorer la communication et diminuer le nombre d'ambiguïtés, l'usage conjoint de plusieurs modes multiplie les problèmes techniques et peut dégrader les performances des usagers.

Dans cet article nous nous intéressons aux entrées des systèmes multimodaux de communication personne-machine. Pour comprendre le but de l'usager, le système doit effectuer correctement la fusion des entrées qui parviennent des différentes modes. Un des points critiques de cette fusion est la résolution des expressions référentielles (ER). Nous proposons un algorithme de résolution de désignations orales aux objets du contexte visuel. Cet algorithme est fondé d'une part, sur des connaissances linguistiques liées à l'expression référentielle utilisée et d'autre part sur la saillance des objets du contexte visuel. Nous pensons que l'attention de l'usager peut être influencée par la présentation des objets dans le contexte visuel (notion de saillance) et nous prenons en compte cette notion dans le processus de résolution des références.

Après une section dans laquelle nous présentons notre contexte de travail et le problème traité, nous exposons les principaux éléments de la solution proposée. Nous commençons par analyser les différents cas possibles à prendre en compte dans la solution puis nous montrons la rôle de la saillance dans le traitement des ERs. Enfin nous détaillons l'algorithme de traitement des ces désignations que nous illustrons par un exemple d'application qui souligne l'intérêt de cet algorithme.

2 Problématique

Notre cadre de travail est le système prototype Géoral tactile (Siroux *et al.*, 1997) qui est implémenté sur la plateforme multiagent DORIS (L'Hour *et al.*, 2004). C'est un système multimodal pour une application de renseignements géographiques et touristiques. L'usager peut demander des informations et la localisation de sites touristiques comme plage, église, camping en précisant un endroit, une zone (dessinée ou située par rapport à un élément géographique ou cartographique particulier : rivière, route, côte) ; il peut également demander la distance et l'itinéraire entre deux localités.

Les modes et modalités mis à la disposition de l'usager dans le système Géoral sont les suivants :

– l'oral en entrée et en sortie du système. L'usager peut formuler ses demandes ou ses réponses aux questions du système par la voix et en langage naturel (LN) de manière spontanée (pas de consignes particulières d'élocution). Certaines réactions du système sont aussi transmises oralement à l'utilisateur.
– le mode visuel : le système affiche sur un écran une carte de la région ; cette carte contient des informations géographiques et touristiques habituelles : routes, rivières, fleuves. Des effets de zoom, de surlignage, de clignotement permettent au système de focaliser l'attention de l'usager.
– le mode gestuel par l'intermédiaire d'un écran tactile : l'usager peut désigner par différents types de geste des éléments sur la présentation affichée à l'écran.

Un dialogue avec Géoral est composé d'un ou plusieurs échanges. Un échange est constitué des tours de communication de l'usager et du système (Bilange, 1992). Un tour de communication de l'usager consiste en un énoncé oral et/ou une entrée gestuelle et celui du système consiste en une sortie orale (par la synthèse de la parole) et un affichage sur l'écran. Par exemple, un échange simple contient deux tours de communication : un tour de l'usager (question) et un tour du système (réponse). Notons qu'un échange peut être imbriqué en contenant d'autres échanges (dans le cas des questions de clarification). Dans un tour de communication d'un usager, le problème pour le système est de résoudre les ERs, c'est-à-dire, trouver le référent d'un symbole dans une modalité en utilisant des informations présentes dans la même ou dans d'autre modalité.

Dans ce cadre, nous avons proposé une nouvelle définition d'un modèle général de traitement des ERs (Choumane & Siroux, 2006). Ce modèle est fondé sur deux principes : la définition de langages associés à chaque modalité (LN, geste, visuel) plus un langage pivot et la création de fonctions reliant certains objets de chaque modalité permettant d'identifier les référents. A chaque tour de communication, les langages encodent les objets issus des modalités et mémorisent aussi certaines parties du contexte courant et de l'historique de l'interaction. Des traitements seront associés à chaque modalité (par exemple : traitement des anaphores pour la LN) et des traitements spécifiques seront mis en place pour la détermination des référents désignés de manière multimodale.

L'objet de l'article est de présenter une partie du traitement du modèle général de résolution des ERs. L'objectif est de trouver l'objet désigné dans le contexte visuel commun entre l'usager et le système (l'écran), et qui représente le référent d'une ER. Ce traitement est lié à deux des langages du modèle général : le langage qui encode le mode oral et celui qui encode le mode visuel.

Il existe plusieurs types d'ER : ERs qui font référence à des entités de l'historique LN comme l'anaphore (Mitkov, 2002), des ERs qui n'ont pas d'antécédents linguistiques parce qu'elles sont employées en première mention (Vieira & Poesio, 2000), (Manuélian, 2003) et/ou qui font référence à des objets dans une autre modalité qui correspond, par exemple, au contexte visuel dans le système Géoral. Ce dernier type d'ER est produit :
– conjointement avec un geste. Dans ce cas, ce sont des ERs à usage déictique dans lequel le référent est l'objet désigné par le geste.
– ou sans geste.

Dans cet article nous nous intéressons uniquement aux ERs en première mention qui sont produites sans geste. Nous discutons dans la suite les circonstances de production de telles ERs et une méthodologie de traitement.

3 Analyse et solution proposée

Nous traitons ici plus particulièrement le cas d'une désignation par l'oral (sans geste) d'un objet du contexte visuel commun entre l'usager et le système (il s'agit de l'écran dans le système Georal). Nous proposons une solution pour les entrées représentées génériquement par un expression régulière « je veux X (le long | à gauche | près de | à l'embouchure |...) de Y » où X $\in$ {les campings, les hôtels, ... } et Y $\in$ {la rivière, la route, ...}. Cependant la portée de cette solution est plus large, en effet, on trouve ce genre d'expressions référentielles dans d'autre tâches multimodales comme dans (Landragin, 2005) et (Qu & Chai, 2006) dans lesquelles on peut déplacer à l'aide de la LN des objets situés.

3.1 Différents cas possibles

Nous pouvons nous trouver face à plusieurs possibilités de référencement illustrées par les exemples 1, 2, et 3.

Exemple 1 je veux les campings le long de la rivière

Exemple 2 je veux les campings le long de la rivière le Léguer

Exemple 3 je veux les campings le long des rivières

Pour l'exemple 1, trois cas de figure sont envisageables :
- il n'existe aucune rivière dans le contexte visuel (pas de résolution). Dans ce cas la réponse à l'usager est une décision de dialogue. Un message d'information et une question de clarification seront suffisants.
- il existe une seule rivière dans le contexte visuel. Dans ce cas le référent de « la rivière » est l'objet sur l'écran qui représente la seule rivière existante.
- il existe n rivières ($n \geq 2$). Dans ce cas nous appliquons une stratégie fondée sur la saillance pour trouver le référent (détails ci-dessous).

Dans l'exemple 2, qui est un cas particulier de l'exemple 1, l'objet désigné est explicitement nommé dans l'énoncé. C'est-à-dire que l'objet désigné par « la rivière » n'apporte pas d'ambiguïté, il s'agit d'une précision par le nom de l'objet affiché sur l'écran qui est la rivière appelée « le Léguer ».

Dans l'exemple 3, l'ER « des rivières » porte la marque du pluriel, elle désigne les objets du contexte visuel de type « rivière ». Il n'y a pas d'ambiguïté et la résolution ne nécessite pas un traitement spécial autre que celui de la détection de l'ER lors de l'analyse syntaxique.

Les exemples 1, 2 et 3 [2] illustrent des débuts de dialogue dans lesquels les ERs sont employées en première mention. Au cours de dialogue, toute ER doit être évaluée en terme de coréférence (reprise) ou de première mention. Pour cette détermination on utilise les liens lexicaux (synonymie, hypernoymie, et méronymie) ainsi que la liste des items lexicaux représentant les objets sur l'écran. Notons que nous traitons les ERs dans un cadre applicatif précis qui permet de restreindre certains traitements.

La question qui se pose est pourquoi un usager est capable de demander des informations par une entrée multimodale aussi « vague » (exemple 1) et sans geste ? Il faut d'abord noté que

[2] Exemples extraits d'une expérimentation légère avec Georal (Siroux *et al.*, 1997).

malgré l'absence d'un geste conjointement à la parole, nous classons cette entrée comme multimodale. En effet, l'usager fait référence à un objet dans une modalité (LN dans l'exemple 1) en s'appuyant sur des informations présentes dans une autre modalité (les objets du contexte visuel dans l'exemple 1). D'une part, dans la production d'une entrée multimodale comme dans l'exemple 1, l'usager s'est appuyé sur le contexte visuel pour désigner son objet parce que cet objet est « suffisamment saillant », pour l'usager, que ce dernier n'a pas complété son entrée par un geste (notons aussi que l'usager peut être confronté à des problèmes de performance : crainte de toucher l'écran, ...). D'autre part, le système prépare et affiche le contexte visuel en fonction de l'importance applicative des objets et calcule donc leur saillance pour cet affichage.

3.2 Notations

La formalisation dans les prochaines sections est fondée sur les notations suivantes :

CVC_c = contexte visuel commun courant entre le système et l'usager.
e un échange donné entre l'usager et le système.
t un tour de communication donné. t de l'usager comprend normalement un énoncé oral et/ou une entrée gestuelle. Dans cet article, il s'agit uniquement d'un énoncé oral.

R={r_k, $1 \leq k \leq K$/ ER(r_k)}, le(s) ER(s) prononcée(s) par l'usager au tour de communication t.

O_k={o_j, $1 \leq j \leq J$}, l'ensemble des objets candidats référents de l'expression référentielle r_k.

$S_c(o_j)$ est la saillance de l'objet o_j dans le contexte visuel courant c (CVC_c).

Pour simplifier la présentation de l'algorithme (cf. section 3.4), nous supposons que dans un tour donné t d'un usager, |R| = 1 (K=1). C'est-à-dire que t ne contient qu'une seule ER qui désigne un objet dans CVC_c. Donc :
– r_1 est la seule ER dans le tour de communication t. Dans l'exemple 1, r_1 est « la rivière ».
– O_1 contient les objets candidats référents de l'expression référentielle r_1.

3.3 Saillance et son utilisation

La saillance intervient fortement lors de l'interprétation d'un énoncé en situation de dialogue ou lors de la compréhension d'un texte : en mettant en avant un élément, elle dirige l'attention sur cet élément et rend sa prise en compte prioritaire dans le processus de résolution des références et des coréférences (Landragin, 2005). Nous trouvons dans la littérature deux types de saillance : la saillance linguistique et la saillance visuelle.

La saillance linguistique, qui dépend uniquement de la modalité LN, constitue, par exemple, une aide à la résolution des anaphores (Lappin & Leass, 1994). Dans toute communication homme-machine dont le mode visuel fait partie, la saillance visuelle constitue un critère d'identification de l'objet désigné et perçu de manière prioritaire (Landragin, 2005).

Notre approche est fondée sur ce que nous appelons la « saillance contextuelle ». Nous visons ainsi une notion plus large que celle de la saillance visuelle. En effet, la saillance contextuelle d'un objet sera affectée durant l'interaction selon que l'objet est désigné ou non par l'usager tout en prenons en compte les caractéristiques visuelles des objets qui peuvent capter son attention. Désormais, nous faisons référence à la saillance contextuelle par le mot « saillance » uniquement. Nous allons montrer, dans la suite de cet article, comment cette saillance permet la

résolution d'une ER sans antécédent linguistique et sans geste accompagnant l'énoncé oral.

Nous distinguons deux moments d'utilisation de la saillance :
– au début d'un dialogue nous attribuons des valeurs par défaut aux objets du CVC. Cette attribution n'est pas nécessairement équiprobable, les valeurs peuvent être en effet liées à l'application. La détermination d'une méthode de calcul numérique de la saillance (Landragin, 2005) ne fait pas partie de cet article. Notons simplement l'existence de plusieurs facteurs qui contribuent à rendre saillant un objet et qui interviennent donc pour la quantification de la saillance de cet objet. Ces facteurs pourront être la couleur, la taille, la complexité, etc. d'un objet. Cette phase d'initialisation sera appliquée à chaque début de dialogue (au moins un échange).
– durant l'interaction, on modifie les saillances des objets qui sont dans O_1 à la fin de chaque interprétation de l'énoncé de l'usager. Ainsi la (les) saillance(s) de(s) référent(s) augmente(nt) et la (les) saillance(s) de(s) autre(s) objet(s) de O_1 diminue(nt). Nous rappelons que l'ensemble O_1 d'un tour de communication d'un usager contient l'objet (les objets) candidat(s) référents de r_1. Soit $S_c(o_j)$ la saillance de l'objet o_j dans le CVC_c. A la fin de l'interprétation de l'énoncé de l'usager, et après avoir utilisé les saillances de CVC_c, nous allons modifier les saillances des objets de O_1. L'interprétation par le système de tout énoncé de l'usager va prendre en compte l'ensemble des informations contextuelles (visuelles, linguistiques, ...).

Voici l'algorithme simplifié de distribution de saillance (a et b sont des constantes à régler, avec $a > 0$ et $b \leq 0$) :

si c'est un début de dialogue **alors**
  pour tout $o_j \in CVC_c$ **faire**
    $S_c(o_j) \leftarrow S_0(o_j)$ (initialisation des saillances)
  fin pour
sinon {c'est la fin de l'interprétation d'un énoncé de l'usager et nous disposons d'un ensemble O_1. Nous allons modifier les saillances des objets de O_1}
  pour tout $o_j \in O_1$ **faire**
    si o_j est un référent **alors**
      $S_c(o_j) \leftarrow S_c(o_j) + a$
    sinon {c'est-à-dire que c'est un objet à pénaliser}
      $S_c(o_j) \leftarrow S_c(o_j) + b$
    finsi
  fin pour
finsi

Cet algorithme est appelé par l'algorithme de recherche du référent ci-dessous (cf. section 3.4). Si l'algorithme de distribution de saillance a un ensemble O_1 en entrée, alors les saillances de ces objets o_j seront modifiées d'une valeur a ou b selon que l'objet est un référent ou non. Les constantes a et b seront réglées lors d'expériences ultérieures à mener.

3.4 L'algorithme de recherche de référent

L'algorithme que nous proposons (organigramme de la figure 1) est constitué de plusieurs phases :

– La première phase est celle de détermination de l'ensemble O_1 des objets candidats référents de r_1. Le critère de candidature est le type de l'objet. Dans l'exemple 1, O_1 est l'ensemble

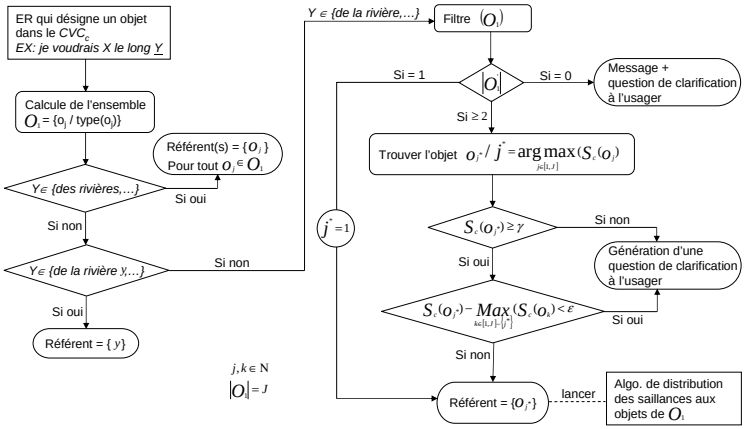


FIG. 1 – Algorithme de recherche de référent, dans le contexte visuel, désigné par l'oral

des objets de type « rivière » et de tous les objets dont leur type appartient à la classe des relations lexicales de « rivière » comme les synonymes, les hypernoymes, et les méronymes. Ces connaissances de synonymie, d'hypernoymie, et de méronymie peuvent être trouvées dans une base de données lexicales comme WordNet lors de l'analyse syntaxique de l'énoncé. Donc O_1 contient les objets de types « rivière », « fleuve », « ruisseau », etc. Pour prendre en compte des éventuels problèmes liés aux performances de l'usager (le cas d'un usager qui prononce « rivière » pour désigner un « ruisseau », ...), nous avons choisi de déterminer O_1 selon la classe lexicale entière. Cependant d'autres choix sont possibles :

1. Dans le cas d'objets qui ne sont pas explicitement typés sur l'écran [3], la détermination de l'ensemble O_1 selon toutes les informations lexicales est nécessaire. Dans l'exemple « je veux les hôtels le long de la route », l'ensemble O_1 contient tous les objets o_j de CVC tel que le type de o_j est égal à « route », « voie », etc.
2. Dans le cas des objets typés explicitement sur l'écran (le cas des rivières dans Géoral) d'une manière efficace qui garantit une meilleur visibilité de ces types, la détermination de l'ensemble O_1 pourrait dépendre uniquement du type de l'ER, d'une partie de sa classe lexicale, etc.

– Dans la deuxième phase nous testons si l'ER en question correspond à un cas particulier tel que ceux des exemples 2 et 3. Dans les deux cas, le(s) référent(s) est (sont) l' (les) objet(s) directement nommé(s) dans l'énoncé oral ou l'ensemble O_1 entier (cf. section 3.1).
– Si l'ER n'est pas un cas particulier (c'est alors l'un des cas de l'exemple 1), nous appliquons un filtre sur O_1. Le critère de filtrage dépend du type des objets. Par exemple, « la qualification » forme le critère de filtrage pour les objets de type rivière. Dans le cas de l'énoncé, « je veux les campings le long de la grande rivière », nous filtrons O_1 en ne choisissant que « les grandes rivières ». Dans le système Géoral, les caractéristiques des objets du CVC comme la taille, la couleur, etc. sont codées en représentation interne.
– Ensuite, nous testons le nombre d'élément de O_1, et trois cas sont à prendre en considération :

[3] Les routes dans le système Géoral ne sont pas étiquetées par « route » sur l'écran, mais par D11, D21, etc. Mais, dans la représentation interne du système, D11, D21, etc. sont codés comme route, voie, etc. Les choix des représentations sur l'écran sont liés à des problèmes de clarté, d'efficacité d'affichage, et d'usage.

1. Si $|O_1| = 0$, alors le message d'information et la question de clarification seront lancés. Si la question est accompagnée d'affichage graphique spécial (zoom, clignotement, etc) les saillances des objets mis en cause par le système seront modifiées. Notons que dans ce cas, il y aura imbrication d'échange.
2. Si $|O_1| = 1$ alors le référent de r_1 est l'objet représenté par o_1.
3. Si $|O_1| \geq 2$ alors nous recherchons l'objet le plus saillant parmi les objets de O_1, c'est l'objet o_{j^*} tel que :

$$j^* = \arg\max_j S_c(o_j)$$

$$avec\ S_c(o_{j^*}) \geq \gamma\ et\ S_c(o_{j^*}) - \max_{k \in [1,J]-\{j^*\}} S_c(o_k) \geq \epsilon$$

γ est un indice de confiance, il est nécessaire pour éviter le choix d'un objet peu saillant. Il dépend de la moyenne des saillances de tous les objets de O_1 et peut être calculé par :

$$\gamma = \frac{\lambda}{J} \sum_{j=1}^{J} S_c(o_j)$$

avec λ est un réel. Nous comptons affiner le calcul de γ par des expériences ultérieures à mener. ϵ est un réel suffisamment grand pour dire que r_1 n'a désigné que l'objet qui a la plus grande saillance o_{j^*}. C'est pour détecter les cas d'ambiguïté.
Si

$$S_c(o_{j^*}) < \gamma\ ou\ S_c(o_{j^*}) - \max_{k \in [1,J]-\{j^*\}} S_c(o_k) < \epsilon$$

Alors le système génère une question à poser à l'usager pour choisir un des objets O_1. Et après la résolution de l'expression référentielle en question, on modifie les saillances des objets de O_1 en faisant appel à l'algorithme de distribution de saillance montré ci-dessus (cf. section 3.3).

Notons qu'après avoir trouvé l'objet désigné, nous devons prendre en compte la préposition prononcée avant Y (Vandeloise, 1986) (« le long de » dans l'exemple 1) pour trouver la partie exactement visée par l'usager comme étant l'espace physique de recherche du thème (les campings dans l'exemple 1). L'influence de cette préposition pourrait être plus intéressante dans le cas de « à l'embouchure de », « à gauche de », etc.

3.5 Scénario d'application

Nous présentons dans cette section un exemple de dialogue entre un usager (U) et le système Georal (S). Les énoncés de U et de S qui sont en italique, correspondent respectivement à des entrées orales et à des sorties orales (par synthèse de la parole). Notons que U_t (S_t) correspond à un énoncé de U (S) au tour de communication numéro t.
Échange $e = 1$
U_1 : *Quel est le nom de ça + geste de pointage sur l'écran*
S_1 : *Bois de Lann ar Waremm*
Dans cet échange, qui se déroule dans un contexte visuel courant CVC_c, l'expression référentielle « ça », à usage déictique, a comme référent l'objet désigné par le geste démonstratif qui accompagne l'énoncé oral. C'est l'objet de CVC_c qui représente le *Bois de Lann ar Waremm*.
Échange $e = 2$
U_1 : *Je veux les campings le long de la rivière*

Le CVC_c de ce tour de communication est issu du contexte visuel précédent après certaine modification dans son contenu comme les saillances des objets. L'ER « la rivière » ne peut pas être résolue linguistiquement, il s'agit d'une première mention qui n'a aucun antécédent dans le discours. Pour cela, nous appliquons l'algorithme proposé ci-dessus pour trouver l'objet référent du CVC_c de l'ER « la rivière ».
r_1 = *la rivière*.
L'ensemble des candidats référents de r_1 est $O_1 = \{o_1, o_2, o_3, o_4\}$, avec o_1 = *Le Léguer*, o_2 = *Ruisseau de Gruguil*, o_3 = *Ruisseau de Kerhuel*, et o_4 = *Rivière des Traouïero*.
Nous remarquons qu'il n'y a pas de qualification de r_1, donc O_1 ne change pas après la phase de filtrage.
Comme $|O_1| = 4$, alors nous allons chercher l'objet o_{j^*}.
Supposons que dans CVC_c nous avons les données suivantes :

$$\begin{cases} S_c(o_1) = 2,\ S_c(o_2) = 1,\ S_c(o_3) = 1,\ et\ S_c(o_4) = 1.5 \\ \lambda = 1 \\ \epsilon = 0.3 \end{cases}$$

alors $j^* = 1$ et $\gamma = 1.375$.
Nous passons maintenant à la phase de validation de l'objet o_1 comme référent. Comme

$$S_c(o_{j^*}) = S_c(o_1) = 2 > \gamma\ et\ S_c(o_{j^*}) - \max_{k \in \{2,3,4\}} S_c(o_k) = S_c(o_1) - S_c(o_4) = 0.5 > \epsilon$$

alors le référent de « la rivière » est l'objet o_1 (le Léguer).
S_1 : *voulez vous les campings le long de la rivière le Léguer ? Merci de confirmer. Si non, veuillez préciser une autre rivière.*
U_2 : *oui*
S_2 : *il y a un camping qui répond à votre requête : camping de Beg Léguer*.

Après la confirmation de l'usager, l'expression référentielle r_1 est considérée, par le système, comme résolue. Ainsi la saillance de l'objet o_1 sera augmentée d'une valeur a. Les autres objets de O_1 (o_2, o_3, et o_4) seront pénalisés en baissant leur saillance d'une valeur b puisqu'ils n'étaient pas désignés.

L'échange numéro 2 n'est pas simple, il contient un sous-échange (S_1, U_2) mené par le système qui pourrait affecter aussi les saillances de O_1. Notons que les saillances des objets du CVC_c seront réinitialisées dès le premier échange du prochain dialogue.

4 Conclusion

Nous avons proposé une solution pour le traitement des désignations orales des objets dans le contexte visuel commun entre l'usager et le système. Cette solution est fondée sur des connaissances sur la modalité langue naturelle, des connaissances sur les objets manipulés, et des connaissances sur les aspects perceptifs (degré de saillance) associés à ces objets. Elle met en valeur un algorithme constitué de plusieurs phases. La phase générale consiste à chercher le référent le plus saillant dans une liste des candidats référents (appelée O_1) avec un indice de confiance et une vérification de validité pour détecter les cas d'ambiguïté. La terminaison du processus de résolution provoque, quelque soit le résultat, la modification des saillances pour le

tour de communication suivant. Nous pensons affiner certains paramètres en menant des expériences avec le système Georal. L'état du module de reconnaissance de la parole de Georal ne permet pas actuellement de mener des expérimentations de validation.

L'apport principal de notre proposition réside dans les possibilités de stratégie de dialogue offerte au système pour faire préciser les référents en cas de problème. En effet, si la liste des objets candidats référents est relativement importante, nous pensons qu'il est fastidieux de proposer à l'usager de choisir un des objets de la liste en lui présentant tous les objets. Le choix de proposer l'objet le plus saillant à l'usager est issu du fait que ce dernier s'est appuyé sur le contexte visuel dans son activité de désignation.

Remerciements

L'auteur remercie Jacques Siroux pour sa collaboration et son aide précieuse pour les travaux exposés.

Références

BILANGE E. (1992). *Dialogue personne-machine. Modélisation et réalisation informatique*. 2-86601-324-7. Hermes.

CHOUMANE A. & SIROUX J. (2006). Toward a generic model including knowledge and treatments for multimodal reference resolution. In V. P. GUERRERO-BOTE, Ed., *Proceedings Inscit2006*, volume 2, p. 298 – 302, Mérida - Spain.

LANDRAGIN F. (2005). Traitement automatique de la saillance. In *Douzième conférence sur le traitement automatique des langues*, p. 263 – 272.

LAPPIN S. & LEASS H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**, 535 – 561.

L'HOUR J., BOËFFARD O., SIROUX J., MICLET L., CHARPENTIER F. & MOUDENC T. (2004). Doris, a multiagent/ip platform for multimodal dialogue applications. *ICSLP*.

MANUÉLIAN H. (2003). *Descriptions définies et démonstratives : analyses de corpus pour la génération de textes*. PhD thesis, Université Nancy 2.

MITKOV R. (2002). *Anaphora Resolution*. 0-582-32505-6. Pearson Education.

QU S. & CHAI J. Y. (2006). Salience modeling based on non-verbal modalities for spoken language understanding. In *ICMI '06 : Proceedings of the 8th International Conference on Multimodal Interfaces*, p. 193–200, New York, NY, USA : ACM Press.

SIROUX J., GUYOMARD M., MULTON F. & RÉMONDEAU C. (1997). Multimodal references in georal tactile. In *Workshop Referring Phenomena In a multimedia Context And Their Computational Treatment, 35th Meeting Of the ACL*, Madrid - Spain.

VANDELOISE C. (1986). *L'espace en français*. Editions du Seuil, Paris.

VIEIRA R. & POESIO M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, **26 (4)**, 539–593.

Combinaison de ressources linguistiques pour l'aide à l'accès lexical : étude de faisabilité

Laurianne SITBON
Laboratoire d'Informatique d'Avignon - Université d'Avignon
Laboratoire Parole et Langage - Université de Provence
laurianne.sitbon@univ-avignon.fr

Résumé. Cet article propose une évaluation combinée et comparative de 5 ressources (descriptive, paradigmatique et syntagmatiques) pour l'aide à l'accès lexical en situation de "mot sur le bout de la langue", en vue de la création d'un outil utilisant la combinaison de ces ressources. En situation de "mot sur le bout de la langue", l'utilisateur n'accède plus au mot qu'il veut dire ou écrire mais est capable d'en produire d'autres sémantiquement associés. L'évaluation se base sur un corpus de 20 mots "sur le bout de la langue" pour lesquels on dispose de 50 groupes de 5 associations sémantiques effectuées par des utilisateurs. Les résultats montrent que les ressources sont complémentaires et peu redondantes. De plus au moins une association proposée parmi les 5 permettrait de retrouver le mot "sur le bout de la langue" dans 79% des cas, à condition de le sélectionner parmi les 2500 mot potentiels. Enfin, les résultats montrent des disparités entre les utilisateurs, ce qui permettrait de définir des profils d'utilisateur pour une amélioration des performances.

Abstract. This paper describes a joint and comparative evaluation of 5 lexical resources (descriptive, paradigmatic and syntagmatic) from a lexical access angle, with the further perspective of constructing a tool based on a combination of these resources to avoid the "tip of the tongue" (TOT) phenomenon. This phenomenon characterises a person who has difficulty in saying or writing an intended word but one who is able to produce semantically associated words. The evaluation corpus is composed of 20 TOT examples each linked to 50 users' association sets of 5 semantically associated words. The results highlight that all the tested resources are complementary. Moreover, 79% of proposed association sets contain at least one association leading to the TOT through its relative words in at least one resource (in the worse case the TOT has to be found among 2500 words). Finally the results show variations between users which could increase performance thanks to user profiles.

Mots-clés : réseaux sémantiques, accès lexical, profil d'utilisateur.

Keywords: semantic networks, lexical access, user profiling.

1 Introduction

Les outils du traitement automatique du langage ont démontré à plusieurs reprises leur capacité à remédier ou compenser des handicaps de langage. Le problème du *mot sur le bout de la langue*, plus généralement connu comme étant un déficit d'accès lexical en production de phrases (Tip

Of the Tongue en anglais (TOT)), est l'une des manifestation de tels handicaps. Si le phénomène a été étudié dans le cadre général (des personnes sans déficit particulier peuvent en souffrir par moments), les observations diffèrent en situation de handicap. En effet les travaux de (Brown & McNeill, 1966) ont montré que en situation de TOT les personnes connaissent des informations sur le mot recherché, d'ordre phonologique aussi bien que sémantique. Les modèles actuels de la production du langage proposent une vision connexioniste (Dell, 1986), où l'information verbale serait stockée dans trois systèmes interconnectés (le système sémantique, le système phonologique et le système orthographique). D'après les travaux de (Burke & Shafto, 2004) sur le déficit d'accès lexical chez les personnes âgées, c'est essentiellement un déficit d'accès phonologique qui est en cause, ce qui signifie que l'accès sémantique est intact. Dans les modèles connexionistes, l'interconnexion entre les unités sémantiques est beaucoup plus importante que la connexion unique entre un mot et sa phonologie ou un mot et sa graphie (qui sont en réalité des connexions des phonèmes vers les graphèmes, plus une connexion du sens vers la graphie qui porte les exceptions), ce qui explique que lorsque les connexions sont "fragilisées" elles sont maintenues au sein du réseau sémantique mais pas pour l'accès phonologique. Les travaux de (Faust & Sharfstein-Friedman, 2003) montrent un comportement similaire chez des adolescents dyslexiques, qui s'explique dans ce cas par un déficit de la conscience phonologique fréquent chez les personnes souffrant de dyslexie.

Un système qui reproduirait le réseau sémantique de chaque individu (et doté d'un convertisseur phonologique robuste entre les sens et les mots) permettrait ainsi de proposer à l'utilisateur en situation de handicap et face à un mot recherché une liste d'hypothèses, obtenue à partir de mots proches dans son réseau sémantique. Ainsi automatiser l'aide à l'accès lexical nécessiterait une représentation exhaustive et individuelle du lexique mental. Dans cet article nous faisons l'hypothèse que la construction d'une telle ressource peut être approchée à l'aide d'une combinaison de ressources déjà disponibles ou réalisables automatiquement, et qui présentent chacune des aspects différents du cheminement de la pensée pour passer de l'idée au mot. Les ressources lexicales auxquelles nous nous sommes intéressé ont déjà fait indépendamment l'objet d'études dans le cadre de l'aide à l'accès lexical (Reuer, 2004). Cependant tous les individus étant différents nous pensons que les ressources lexicales peuvent être utiles à des individus différents à différents degrés. On peut aussi penser que les voies empruntées par le cerveau pour accéder à un mot dépendent de caractéristiques linguistiques ou sémantiques de ce mot. Encore une fois, la ressource la plus appropriée dans un cas précis pourrait dépendre de la nature sémantique ou syntaxique du mot recherché.

L'idée à long terme est d'implémenter ce système avec des associations audio, ce qui permettrait une utilisation en situation quotidienne. Les bonnes performances des systèmes de reconnaissance automatique de la parole nous laissent penser que la prise en compte de peu d'hypothèses de reconnaissance serait suffisante pour converger vers le sens recherché. L'aide à la formulation ou l'enrichissement semi-automatique de requêtes en recherche d'information est également un domaine d'application qui pourra bénéficier de ce système.

La conception d'un tel système soulève plusieurs questions auxquelles nous tenterons de répondre dans cet article. Dans un premier temps le choix des systèmes utilisés est justifié dans la première section. Il doit permettre un recouvrement maximal des types de relations possibles (paradigmatiques, syntagmatiques ou descriptives) ainsi que des domaines sémantiques (journalistique, littéraire, générique). Ensuite l'intérêt d'une telle combinaison doit être attesté à travers une évaluation combinée des systèmes unitaires sélectionnés. Pour cela la seconde section présente l'élaboration d'un corpus d'évaluation avec l'aide de 50 participants qui ont produit des associations sémantiques sur des simulations de TOT sur 20 mots.

2 Les ressources linguistiques testées

Les approches existantes dans le domaine du "mot sur le bout de la langue" ont souvent évoqué l'usage de ressources de natures différentes. (Lortal *et al.*, 2004) proposent d'étendre les classes sémantiques du réseau de relation sémantico-pragmatiques SVELTAN à l'aide d'EuroWordNet[1] et montrent l'intérêt de les utiliser ensemble pour une tâche consistant à retrouver 10 mots supprimés dans des documents. (Zock, 2002) évoque trois accès aux TOT *via* des ressources différentes sans les évaluer : un par la forme du mot, graphique ou phonologique, un autre par son sens en termes de collocations dans le domaine sémantique auquel il appartient, et un accès par la fonction grammaticale. (Reuer, 2004) argumente en faveur de quatre types de ressources linguistiques différentes : des réseaux sémantiques construits manuellement , des réseaux paradigmatiques et syntagmatiques construits automatiquement, et des réseaux phonologiques et morphologiques.

Etant donné que l'accès phonologique est justement la cause du TOT dans les situations de handicap, il n'est pas pertinent dans notre cas d'utiliser des ressources pour l'accès phonologique, et nous nous sommes donc concentré sur la représentativité des 3 premiers types de ressources. Les relations paradigmatiques sont des liens sémantiques de type synonymiques ou hiérarchiques qui peuvent dépendre du contexte. Les relations syntagmatiques sont des associations mnésiques, issues d'un contexte stocké dans la mémoire à long terme. Elles peuvent être en parties extraites à l'aide de cooccurrences. Nous avons ajouté à ces ressources des relations descriptives issues d'un dictionnaire.

2.1 Les relations descriptives

La manière la plus naturelle d'expliquer à une personne ou à un système le mot que l'on recherche est de le décrire avec d'autres mots, même si il s'avère que cette description est parfois impossible car elle implique le même amorçage lexical que celui du mot recherché. Les dictionnaires constituent de bons référentiels en ce qui concerne la description des mots et disposent généralement d'une bonne couverture. Nous avons utilisé un dictionnaire en ligne (`http://fr.answers.com` propose gratuitement un dictionnaire du français) pour nos expériences. Tous les mots d'une définition n'étant pas nécessairement reliés sémantiquement au mot défini (on rencontre beaucoup de mots outils), des mots clés sont extraits en fonction de leur fréquence relative (*tf/idf*) dans un corpus généraliste (extraits du journal *Le Monde*). Les relations sémantiques entre les mots induites dynamiquement par cette technique ne sont pas des relations symétriques (par exemple le mot *pied* peut être dans la définition de *chaussette* mais l'inverse est peu probable).

2.2 Les relations paradigmatiques

EuroWordNet est une base de données multilingue construite sur le modèle de Wordnet [2]. A chaque mot est associé un certain nombre de sens, qui sont hiérarchisés selon des relations paradigmatiques : synonymie, méronymie, hyponymie. La base de données EuroWordNet française dont nous disposons contient ces informations pour les noms, les verbes et les adjectifs.

[1] `http://www.illc.uva.nl/EuroWordNet/`
[2] `http://wordnet.princeton.edu/`

Le principal inconvénient de cette ressource pour l'accès lexical est que les relations ne sont pas quantifiables, on ne peut pas leur associer de score. Le niveau dans la hiérarchie d'un méronyme ou d'un hyponyme n'est pas non plus utilisable car selon les mots, le niveau de détails hiérarchique peut être très variable. Par exemple une *pomme* sera un *fruit* alors qu'un *citron* sera un *agrume* qui est aussi un *fruit*. Pourtant la relation *pomme/fruit* est aussi forte que la relation *citron/fruit*.

2.3 Les relations syntagmatiques apprises automatiquement

Les relations syntagmatiques sont les moins délimitées, et il est impensable d'assurer leur représentativité. Cependant nous nous sommes concentrés sur trois types de corpus de natures différentes : des articles de presse, des romans, et le web. Des méthodes d'extraction de relations différentes sont appliquées sur chacun des corpus, mais elles sont toutes basées sur l'idée de cooccurrences entre les mots associés et le mot recherché.

Un premier réseau de relations syntagmatiques appris automatiquement est une carte sémantique apprise à l'aide de l'outil Infomap[3] sur le corpus littéraire corpatext[4]. Infomap crée des cartes sémantiques en se fondant sur le principe de LSA (Latent Semantic Analysis) (Deerwester *et al.*, 1990), c'est à dire qu'il effectue une réduction de l'espace lexical composé des mots du corpus afin de les regrouper en classes sémantiques, en se basant sur les cooccurrences des mots. (Landauer *et al.*, 1998) a montré la représentativité du lexique mental à travers LSA, qui est utilisé pour cette propriété par des psycholinguistes. Cependant cette approche ne permet pas d'explorer tous les sens d'un mot, au contraire elle le classe dans sa classe sémantique la plus représentée. La création de *contextonymes* proposée par (Ji *et al.*, 2003) permet de distinguer les différents sens en créant des classes de termes associés, mais nous ne les avons pas expérimentés ici.

Un second réseau de relations syntagmatiques est un réseau de cooccurrences. Il est utilisé par (Ferret & Zock, 2006) pour des travaux similaires. Il est construit selon la méthode proposée par (Church & Hanks, 1990) qui calcule l'information mutuelle entre les termes d'une fenêtre glissante de 20 mots sur le corpus journalistique (extraits du journal *Le Monde*). On obtient ainsi des scores de cohésion entre les paires de mots cooccurrents du corpus. Cependant, comme pour les relations paradigmatiques apprises par LSA, les cooccurrences ne résolvent généralement pas les ambiguïtés de sens lorsqu'un sens est plus fréquent qu'un autre.

La troisième ressource créée n'est pas un réseau de relations mais plutôt un générateur dynamique de relations syntagmatiques, selon le même principe que la génération de relations descriptives. Elle permet d'obtenir les mots clés des 10 premières réponses du moteur de recherche Google [5] à la requête correspondant au mot que l'on cherche à mettre en relation. On considère que les mots de plus grande fréquence relative (*tf/idf*) dans les 10 sites les plus pertinents selon Google sont des cooccurents importants et sont relié au mot par une relation syntagmatique.

(Joyce, 2005) propose la construction d'un réseau de relations syntagmatiques à partir d'un réseau d'associations qui reposerait sur l'expérience personnelle des individus, en demandant à beaucoup de personnes de participer à la création de ce réseau. Il s'agit d'une partie d'un programme de création de ressources à très grande échelle qui est pratiqué sur la langue japonaise.

[3] http://infomap-nlp.sourceforge.net/
[4] http://www.lexique.org/public/corpatext.php
[5] http://www.google.fr

3 Constitution d'un corpus d'associations sémantiques

Jusqu'à ce jour les réseaux proposés ont été évalués dans des cadres d'autres applications, comme la traduction ou l'aide à la rédaction. Nous proposons un cadre évaluatif reflétant la tâche d'un système d'aide à l'accès lexical en situation quotidienne pour des personnes âgées ou dyslexiques, en composant un réseau d'associations réelles à l'aide de participants.

Pour 20 mots sélectionnés (TOT), 50 personnes ont donné les 5 premiers mots qui leur venaient à l'esprit en association avec le mot proposé, en évitant toute forme de digression. Cette expérience bénévole a été réalisée à l'aide d'une interface web, de manière à laisser les utilisateurs libres de leurs conditions d'expérimentation. La consigne exacte était : *"Cette expérience s'intéresse aux associations d'idées. Il va vous être proposé une série de 20 mots auxquels vous devrez associer à chaque fois les 5 premier mots qui vous viennent à l'esprit. par exemple : OREILLER : plumes, taie, bataille, dormir, moelleux. Ces 5 mots ne doivent pas nécessairement être reliés entre eux, mais doivent tous être associés au mot de départ. Ainsi, on n'attend PAS pour oreiller : plumes, oiseau, arbre, feuille, papier ... (ici chaque mot est relié au précedemment cité)."*

Les 20 TOT ont été sélectionnés de manière à ce qu'ils fassent tous partie des ressources linguistiques statiques (EuroWordNet, la carte LSA sur Corpatext et le réseau de cooccurrences). La liste des mots ainsi que leurs fréquences dans Lexique 3 [6] (fréquence dans des dialogues issus de films, fréquence dans un corpus de livres) est dans le tableau 1. La plupart sont des noms communs, mais il y a aussi des adjectifs et des noms propres, ainsi que des noms d'origine étrangère. Plusieurs de ces mots appartiennent à plusieurs catégories syntaxiques (java est à la fois un nom propre et un nom commun, massif est à la fois un adjectif et un nom commun), certains sont des noms communs repris par des marques commerciales (Casino, Lion), et d'autres sont aussi fortement polysémiques (facteur, baril, brioche, quiche). Tous les mots sont de fréquence inférieure à 20, donc peu fréquents voir rares.

TOT	f. film	f. livre	mot	f. film	f. livre	mot	f. film	f. livre
veau	6,20	16,96	baril	4,22	3,04	casino	17,41	10,81
entretien	17,71	27,77	cambriolage	9,34	2,77	java	0,72	2,30
menhir	0,18	0,68	quiche	0,66	0,68	kleenex	2,11	2,91
lion	17,05	33,04	brioche	7,29	7,09	facteur	11,27	14,32
festin	5,12	5,68	chaussette	14,58	22,84	massif	8,13	22,30
virtuel	2,53	2,16	rugby	1,87	3,11	landau	1,20	4,59
snowboard	N/C	N/C	oie	5,90	9,32			

TAB. 1 – Mots sélectionnés pour la constitution du corpus : chaque utilisateur devait proposer les 5 premiers mots ou expressions qui leur venaient à l'esprit. Les fréquences de chaque mot dans des dialogues de films (f. film) ou dans un corpus de livres (f. livre) montrent qu'il s'agit de mots peu fréquents dans le langage oral comme dans le langage écrit.

Les 50 personnes sont des adultes d'âges, de professions et de niveaux d'études variés, on part ainsi sans a priori social sur l'échantillon de population qu'elles représentent. Elles ont toutes proposé les 5 associations demandées. On note que des participants (non retenus dans le corpus) n'ont pas pu terminer l'expérience pour qui il était trop difficile de trouver plus d'un mot associé.

Le corpus ainsi créé contient pour chacun des 20 TOT au total 250 associations. En moyenne, on dénombre 76,8 associations différentes par TOT. [7]

[6] `http://www.lexique.org`

[7] Le corpus est disponible sur demande par mail à l'auteur.

4 Etude numérique

On cherche à répondre à l'aide du corpus à trois grandes questions, à savoir : est-il utile de combiner plusieurs systèmes ? Y-a-t-il des mots pour lesquels certaines ressources sont plus représentatives du lexique mental que d'autres ? Est-il possible de déterminer des profils d'utilisateurs dans la prise en compte des ressources ?

Pour tenter de répondre à ces questions, nous avons recherché le TOT dans les 5 environnements sémantiques (listes de mots relatives à chaque ressource) de chaque association proposée pour ce TOT (soit 250 par TOT). La figure 1 illustre le principe de cette évaluation. Pour chaque association, on recherche les mots en relation dans chacune des 5 ressources, en limitant aux 100 premiers mots (au sens de scores lorsqu'ils sont disponibles). Le TOT est considéré retrouvé si fait partie d'une relation retrouvée par une ressource. En effet les relations sont des termes ou des expressions nominales. Ainsi si le mot recherché est *tarte*, l'expression *tarte à la fraise* sera acceptable.

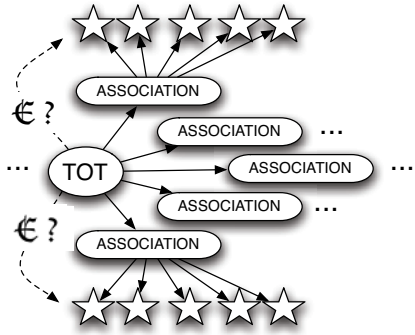


FIG. 1 – Evaluation des ressources : on recherche le TOT dans les listes de mots issues des relations de chacun des mots associé à ce TOT par un utilisateur dans chacune des 5 ressources disponibles.

Ce mode d'évaluation recherche le TOT dans 5×5 listes de 100 mots, soit une liste de 2500 mots différents au maximum. Sachant qu'au final on souhaite retrouver le TOT dans une liste de 10 mots proposés à l'utilisateur en fonction des associations qu'il propose, il ne parait pas pertinent d'augmenter le nombre de mots contenus dans les listes retournées par chaque ressource. On définit les associations "utiles" (AU) comme les associations proposées par les utilisateurs et pour lesquelles la liste de mots (obtenue à partir de relations dans au moins l'une des 5 ressources (500 mots) ou bien dans une ressource en particulier (100 mots) contient le TOT.

4.1 Couverture entre les ressources

La couverture de l'ensemble des ressources est estimée à partir du nombre d'associations comprenant le TOT dans les relations d'une seule des 5 ressources disponibles. On calcule le pourcentage de ces associations par rapport au nombre total d'associations comprenant la cible dans au moins une des 5 ressources.

relations	descriptives	paradigmatiques	syntagmatiques		
	def	ewn	lsa	coocc	web
%ressource	40%	33%	23%	47%	70%
%global	6%	6%	2%	16%	44%

TAB. 2 – Pourcentages de rappel unique (nombres d'associations pour lesquelles la ressource est la seule à retrouver le TOT dans les mots reliés), soit par rapport au rappel global pour cette ressource (nombre d'associations pour lesquelles la ressource retrouve le TOT dans les mots reliés) (%ressource) , soit par rapport au nombre total d'AU (associations "utiles") (%global).

Les pourcentages de rappel unique de chaque ressource par rapport à son rappel total, consignés dans le tableau 2, sont compris entre 23% et 70%, ce qui signifie que chacune des ressources est essentielle pour une bonne part des informations auxquelles elle permet d'accéder, étant donné que dans ces cas là elle est la seule à pouvoir proposer le TOT recherché à l'aide de l'AU. Les pourcentages de rappel unique par rapport à la globalité des réponses montrent néanmoins que l'impact des ressources descriptives ou paradigmatiques est beaucoup moins important dans l'ensemble que les relations syntagmatiques.

4.2 Résultats par mots

Nous nous sommes penchés sur la répartition des relations apportées par chacune des ressources pour chaque TOT du corpus. Pour cela nous avons observé les cas où les associations proposées par les utilisateurs contenaient le TOT dans chacune des ressources. Les résultats dans le tableau 3 montrent tout d'abord que certains mots, comme *veau*, *entretien* ou *rugby* sont plus enclins à être retrouvés par l'ensemble des ressources, alors que d'autres comme *cambriolage* ou *menhir* nécessite la complémentarité de deux ressources (en effet dans ces deux cas la somme des associations reliant au TOT pour chaque ressource est pratiquement égale au total d'associations reliant au TOT, ce qui signifie que ce sont des associations différentes qui ont permis de retrouver le TOT pour chaque ressource). Par ailleurs certains mot ne sont atteignables que par une seule ressource, et pas la même dans tous les cas. Ainsi, *snowboard* n'a été retrouvé que par des associations de relations dans la définition, facteur n'a été atteint pratiquement que à l'aide du réseau de cooccurrences, *festin* majoritairement par la carte sémantique LSA sur le corpus littéraire.

Cependant la couverture globale reste assez faible (33 % des associations permettent de retrouver le TOT pour lequel elles ont été citées). Ainsi nous nous sommes interrogés sur les cas où au moins une ou deux, où toutes les associations, et où la première association donnée par un utilisateur est une AU. On dénombre ainsi les groupes d'associations (5 associations proposées par un utilisateurs pour un TOT) où N associations sont "utiles". Ainsi le tableau 3 contient les pourcentages de ces cas pour chaque TOT. Le pourcentage de cas où les 5 associations sont des AU est de 0% dans la plupart des cas, ce qui exclut toute possibilité d'intersection entre les différents réseaux de l'ensemble des associations proposées pour une combinaison des ressources. Par ailleurs pour 80 % des TOT plus de 70% des groupes d'associations contiennent au moins une AU. Les pourcentages de cas où la première association est "utile" sont à considérer en regard des cas où au moins une association est "utile". On constate alors que si pour des mots comme *landau* ou *veau* la première association est presque toujours une AU, pour des mots comme *lion* ou *java* ou *entretien* l'écart est important ce qui signifie que les AU viennent plutôt après la première.

Mot cible	% d'AU						% cibles : N AU			% cibles :
	def	ewn	lsa	coocc	web	total	N > 0	N > 1	N = 5	1ere AU
veau	23,6	43,2	27,2	46,8	54	80,8	100	98	32	96
baril	0,8	0	0	26,8	24	36	86	62	0	60
casino	0	0	0	32	13,6	44,8	100	76	0	70
entretien	10	10,4	10	0	23,2	23,6	84	30	0	36
cambriolage	0,8	12	0	0,4	11,6	24	72	38	0	40
java	0,4	0,8	0	0,4	4	4,8	24	0	0	0
menhir	13,6	0	0	0	33,2	46,8	96	82	4	62
quiche	0	0	0	0	34,4	34,4	90	64	0	74
kleenex	0	0	0	0	15,2	15,2	72	2	0	50
lion	14,8	4	3,2	0,8	16	30	88	50	0	36
brioche	0,4	0,4	0	0	39,6	40	94	70	0	52
facteur	1,2	1,2	0	13,6	2,4	17,6	74	12	0	28
festin	2,8	0,8	12,4	0	5,6	21,6	68	34	0	48
chaussette	1,6	4,8	0	0	6,8	10	40	8	0	4
massif	1,2	15,6	0	19,2	4,8	38,8	92	68	2	52
virtuel	0	0	0	17,2	24	28,8	82	46	0	30
rugby	17,2	10,4	0	46,8	20,8	63,6	98	90	14	84
landau	0,4	10,4	0	0	29,6	30	98	50	0	84
snowboard	5,2	0	0	0	0	5,2	26	0	0	6
oie	10,4	16,4	18	31,6	48,4	68,4	100	98	8	78
Moyenne	5,22	6,52	3,54	11,78	20,56	33,22	79,2	48,9	3	49,5

TAB. 3 – Pourcentages d'associations "utiles" (AU), parmi les 250 proposées pour chaque mot. Les ressources sont les mots clés du web (web), les mots clés des définitions (def), les liens dans EuroWordNet (ewn), les liens dans la carte sémantique (lsa), et les liens dans le réseau de cooccurrences (coocc). Pourcentages de TOT avec au moins une AU, au moins 2 AU, ou exactement 5 AU, et pour lesquels la première association proposée est "utile".

4.3 Profils d'utilisateurs

Nous faisons l'hypothèse que chaque personne utilise des voies d'accès privilégiées en fonction d'un profil cognitif, et que chacune de ces voies correspond à une des cinq ressources proposées. Nous supposons que s'il est possible de catégoriser automatiquement les données de chaque personne, c'est que d'une part il existe des profils reconnaissables et d'autre part on doit les prendre en compte.

Afin de déterminer si l'on pouvait dégager des profils d'utilisateurs, nous avons effectué une catégorisation automatique des répartitions d'AU pour chaque ressource par utilisateurs en un nombre de classes variable à l'aide de l'algorithme K-Means implémenté dans la plate-forme WEKA. On a ainsi pu obtenir dans la meilleure configuration trois classes de répartitions, dont deux majoritaires qui se différencient essentiellement par la capacité de la ressource de mots clés du web à retourner le TOT à partir des associations proposées. La répartition des AU pour les 20 TOT de chaque utilisateur montre qu'on peut les séparer en deux classes. En effet pour chaque utilisateur une majorité de ses associations appartient à une des deux classes (33 utilisateurs dans la classe favorisant les mots clés web, et 17 dans l'autre).

Cependant, la variété des associations et ressources "utiles" est également fortement liée aux TOT eux mêmes, comme le montrent les résultats du tableau 3. Nous avons donc exploré une autre méthode afin de définir s'il existe ou non des profils d'utilisateurs. Pour cela, nous avons appliqué l'algorithme EM (Expectation-Maximization) sur les données comprenant pour chaque utilisateur et chaque ressource le nombre total d'associations "utiles" parmi les 100 proposées pour les 20 mots réunis. On définit 5 paramètres, qui correspondent aux 5 ressources. Le tableau 4 décrit la composition des 3 clusters proposés par l'algorithme, à travers les moyennes

ressource :	Cluster 1		Cluster 2		Cluster 3	
attribut	Moy.	E.C.	Moy.	E.C.	Moy.	E.C.
web	20,6	3,1	25,2	2,3	13,6	2,6
def	4,6	1,4	6,9	2,2	4,3	1,6
ewn	5,9	1,9	9,4	1,6	3,7	1,3
lsa	3,7	1,1	4,5	1,4	1,8	1
coocc	11,6	1,8	13,4	2,9	9,7	2
instances	48%		32%		20%	

TAB. 4 – Clusters proposées par l'algorithme Expectation-Maximization : valeurs moyennes et écart type pour chaque attribut (pourcentage d'associations utiles par individu pour l'ensemble des 20 mots, pour chaque ressource), et nombre d'instances attribuées à chaque cluster.

pour chaque attribut. L'écart type indique la consistance des données autour de ces moyennes.

Les clusters décrits dans le tableau 4 montrent trois aspects de la prise en compte des ressources pour un profil d'utilisateur. Tout d'abord le cluster 2 se distingue par une place plus importante à apporter aux ressources descriptives et paradigmatiques telles que les définitions et EuroWordNet, alors que les individus du cluster 1 seront majoritairement aidés par les mots clés du web et les cooccurrences. Les individus du cluster 3 se caractérisent essentiellement par leur incompatibilité avec l'ensemble des ressources sélectionnées, c'est à dire que les scores sont faibles pour toutes les ressources. Cependant, il faut noter que de ce fait les ressources sont à prendre en compte pratiquement sur un pied d'égalité. Dans tous les cas cependant, la part accordée aux associations reliées par les mots clés sur le web sont à prendre en compte en large majorité.

5 Conclusions et perspectives

L'évaluation combinée des ressources proposées pour l'aide à l'accès lexical dans le cadre de la recherche du TOT dans des listes de mots obtenues à partir d'associations montre qu'elles sont fortement complémentaires. Cependant le taux de rappel global (nombre d'AU par rapport au nombre total d'associations proposées) n'est que de 33%, ce qui implique qu'une combinaison des ressources devra être capable de sélectionner les associations utiles, puis en extraire le TOT. Dans 79% des exemples du corpus les TOT sont accessibles via au moins une AU, ce qui est la limite maximum du rappel que pourra avoir un système combinant les ressources sélectionnées.

Les relations syntagmatiques obtenues à l'aide des premières pages retournées par Google sont les plus riches en termes de rappel global (21%), ce qui ne signifie pas nécessairement que la technique d'extraction est performante mais plutôt que le corpus représenté par les pages du Web est peut être plus pertinent. De la même manière, le faible rappel de la carte sémantique apprise à l'aide de LSA (3,54%) ne remet absolument pas en cause la qualité de la méthodologie, mais doit être imputé à la nature du corpus à partir duquel la ressource est construite. Nous n'avons pas encore évalué les méthodologies d'extraction de relations syntagmatiques, et cela pourra être intéressant afin de sélectionner la représentation la plus adaptée à chaque nature de corpus, et ainsi améliorer le rappel global.

D'autre part, les associations proposées n'ont pas été caractérisées d'un point de vue psychologique, mais intuitivement on remarque des tendances. Par exemple, beaucoup d'associations sont des noms de couleurs, ce qui suggère des associations sémantiques visuelles plus que verbales. De la même manière 18 personnes ont associé le mot chien au mot facteur, ce qui correspond plus à une image de carte postale ou de film (où le facteur se fait mordre) qu'à un

réel lien sémantique.

Le principal problème qui reste en suspens est comment peut-on pondérer les listes de mots issues de chaque réseau pour parvenir à une liste unique ordonnée, sachant qu'on ne peut présenter qu'une dizaine de mots ou expressions à l'utilisateur pour rester raisonnable en termes d'ergonomie. Au delà de scores de pertinence ou de caractérisation des relations disponibles dans les ressources, les mots concernés par le TOT étant essentiellement des mots peu fréquents, on peut pondérer les mots proposés en fonction de leur fréquence. Etant donné la possibilité de définir des profils utilisateurs, il est peut être également possible de pondérer les mots proposés en fonction de la ressource dont ils sont issus, et d'apprendre ces paramètre de manière empirique. La pondération doit aussi prendre en compte l'appartenance du mot proposé à la sémantique globale dégagée par les associations proposées, ce qui reviendrait à désambiguïser les associations.

Références

BROWN R. & MCNEILL D. (1966). The tip of the tongue phenomenon. *Journal of verbal learning and verbal behaviour*, **5**, 325–337.

BURKE D. M. & SHAFTO M. A. (2004). Aging and language production. *American psuchological society*, **13**(1), 21–24.

CHURCH K. & HANKS P. (1990). Word association norms, mutual information , and lexicography. *Computational Linguistics*, **16**(1), 177–210.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**(6), 391–407.

DELL G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, **93**, 283–321.

FAUST M. & SHARFSTEIN-FRIEDMAN S. (2003). Naming difficulties in adolescents with dyslexia : application of the tip-of-the-tongue paradigm. *Brain and cognition*, **53**(2), 211–217.

FERRET O. & ZOCK M. (2006). Enhancing electronic dictionaries with an index based on associations. In *Coling/ACL joint conference*, Sydney, Australia.

JI H., PLOUX S. & WEHRLI . E. (2003). Lexical knowledge representation with contextonyms. In *MT Summit IX*, New Orleans, USA.

JOYCE T. (2005). Constructing a large-scale database of japanese word associations. *Glottometrics*, **10**, 82–98.

LANDAUER T., FOLTZ P. W. & LAHAM . D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, **25**, 259–284.

LORTAL G., GRAU B. & ZOCK M. (2004). Système d'aide à l'accès lexical : trouver le mot qu'on a sur le bout de la langue. In *TALN'04*, p. 259–268.

REUER V. (2004). Language resources for a network-based dictionary. In *Workshop on Enhancing and Using Electronic Dictionaries ; following COLING 2004*, p. 81–84.

ZOCK M. (2002). Sorry, but what was your name again, or, how to overcome the tip-of-the tongue problem with the help of a computer ? In *COLING-Workshop on building and using semantic networks*, Taipei, Taiwan.

Index par auteurs