

Une approche linguistique et statistique pour l'analyse de l'information en corpus

Yannick Toussaint¹, Fiammetta Namer²,
Béatrice Daille³, Christian Jacquemin⁴,
Jean Royauté⁵ et Nabil Hathout⁶.

¹LORIA-INRIA UMR 7503, Nancy, ²Université de Nancy II,
³IRIN, Nantes, ⁴LIMSI, Orsay, ⁵INIST-CNRS, Nancy, ⁶INaLF-CNRS, Nancy.

Résumé

Cet article présente une chaîne de traitement automatique réalisée dans le cadre du projet ILIAD (Informatique Linguistique et Infométrie pour l'Analyse de grands fonds Documentaires) du GIS Sciences de la Cognition. Cette chaîne est dédiée à l'analyse de l'information à partir de corpus de textes de très grand volume, en français. Elle est expérimentée sur un corpus de 2,5 Mb et a conduit à la création de 50 classes de termes. Ces classes sont construites sur la base de la cooccurrence des termes et représentent des connaissances du domaine. Les différentes étapes de la chaîne associent des méthodes linguistiques informatiques et des méthodes statistiques : pré-traitement des textes, étiquetage, morphologie, terminologie et analyse des documents. Pour chacune d'entre elles, nous présentons les méthodes, les outils ainsi que leur évaluation.

1. Introduction

Cet article présente ILIAD, une chaîne de traitement automatique pour l'analyse de l'information contenue dans des corpus de textes de grande taille. Cette chaîne de traitement associe des méthodes informatiques linguistiques et des méthodes statistiques, assurant, pour chacune des étapes, une grande robustesse. De plus, son architecture ouverte facilite la prise en compte de nouvelles fonctionnalités par l'insertion de nouveaux modules de traitement.

L'analyse de l'information peut être définie comme un ensemble d'outils et de méthodes permettant à un opérateur humain de collecter l'information contenue dans un corpus sans le lire de façon séquentielle. Cet opérateur peut être ainsi capable d'analyser l'avancement d'un domaine scientifique ou technique en considérant simplement un ensemble de classes de termes, construites à partir du corpus initial.

ILIAD a été expérimenté sur un corpus de 2,5 Mb de textes en français sur le domaine de l'agriculture. Ce sont des résumés de notices bibliographiques et constituent de ce fait, un corpus d'une grande homogénéité. L'analyse par ILIAD produit 50 classes actuellement utilisées par les experts pour la veille technologique ou la recherche d'information.

Après une présentation de l'architecture globale du système, nous décrirons les différentes étapes du traitement, les méthodologies adoptées (basées sur les avancées récentes de la termi-

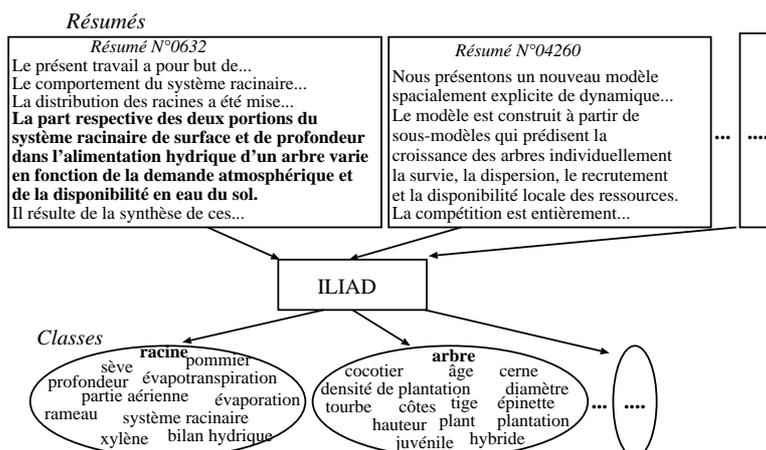


FIG. 1 – Du texte à la classification

nologie, tant du point de vue linguistique que du point de vue de la représentation des connaissances) puis leur évaluation.

2. Architecture

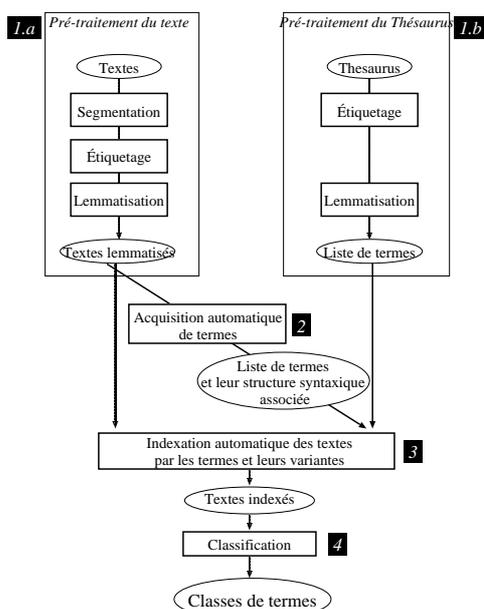


FIG. 2 – Architecture globale

La Figure 1 donne un exemple de classes construites à partir d'un ensemble de résumés. Chaque classe regroupe, sur la base d'une mesure de similarité, les termes fréquemment cooccurents dans les textes. Ce processus est divisé en quatre étapes (voir Figure 2): (1.a) et (1.b) pour le pré-traitement des textes et du thésaurus (étiquetage morpho-syntaxique, segmentation et lemmatisation), puis (2) pour l'acquisition des termes, (3) pour l'indexation, et (4) pour la classification.

Chaque étape repose sur des outils ou méthodes existants (ACABIT, FASTER, SDOC) qui sont décrits dans les sections suivantes. ILIAD a été appliqué à un corpus de 7272 résumés de textes en français (2,5 Mb) sur le domaine de l'agriculture, extraits de la base PASCAL¹ (corpus désigné par [AGR]), et utilise le thésaurus AGROVOC² qui est une taxinomie d'environ 15 000 termes, associés à leurs synonymes et enregistrés dans un format

proche de la norme SGML.

1. PASCAL est la base documentaire scientifique développée et maintenue par l'INIST-CNRS, France

2. Thésaurus multilingue développé par AGRIS - FAO (Unité de traitement AGRIS ; Organisation des Nations Unies pour l'alimentation et l'Agriculture).

3. Des textes à la classification

Nous illustrerons la chaîne complète d'analyse par la phrase exemple (\mathcal{P}), extraite de notre corpus (comme le montre la Figure 1):

Phrase (\mathcal{P}):

La part respective des deux portions du système racinaire de surface et de profondeur dans l'alimentation hydrique d'un arbre varie en fonction de la demande atmosphérique et de la disponibilité de l'eau du sol.

3.1. Préparation des Corpus

Les textes et le thésaurus sont préparés par deux processus identiques, c'est pourquoi nous ne détaillons que celui concernant les textes.

Ceux-ci sont tout d'abord segmentés en phrases, et mis dans un format proche de SGML, puis l'étiqueteur de Brill³(1993) associe à chaque mot sa catégorie grammaticale la plus probable. L'application de ces deux tâches sur la phrase (\mathcal{P}) du § 3 produit (\mathcal{P}_{tag}):

(\mathcal{P}_{tag}): < txt id = 006232 >< ph id = 9 >La/DTN:sg part/SBC:sg respective/ADJ:sg des/DTC:pl deux/CAR portions/SBC:pl du/DTC:sg système/SBC:sg racinaire/ADJ:sg ... varie/VCJ:sg ... < /ph >< /txt >

où DTN est la catégorie des déterminants, SBC des noms, DTC des prépositions contractées, ADJ des adjectifs, CAR des cardinaux, et VCJ des verbes conjugués.

La dernière étape dans la préparation du texte est la lemmatisation des mots fléchis, au moyen d'un programme basé sur des règles, qui associe à chaque forme fléchie un lemme et un ensemble de traits flexionnels.

La morphologie flexionnelle du français est complexe (en comparaison, par exemple, avec celle de l'anglais); on compte environ 80 modèles de conjugaison et autant de terminaisons pour les verbes (cf. (Bescherelle, 1990)), et plus de 40 familles flexionnelles pour les noms et adjectifs. De nombreuses terminaisons sont ambiguës (par exemple "e" ou "s")⁴ ce qui fait que leur interprétation dépend de la catégorie du mot auquel elles sont attachées.

De nombreux lemmatiseurs du français, utilisés en recherche d'information sont, pour cette raison, basés sur la consultation d'un dictionnaire (cf. (Savoy, 1993)). D'autres systèmes, basés, au moins partiellement, sur règles (cf. (Karttunen, 1994), (Guilbaud & Boitet, 1997)), ne sont pas directement comparables avec celui-ci, car ils exploitent un texte non-étiqueté. Par conséquent, le programme d'ILIAD, qui prend en entrée le texte étiqueté, ne doit pas se préoccuper de désambiguïsation. Ne dépendant pas d'un dictionnaire (sauf pour quelques listes d'exceptions aux règles utilisées), il est capable de prendre en compte les mots inconnus et les néologismes qui sont supposés avoir un comportement flexionnel régulier. Enfin, de par sa conception, le lemmatiseur est capable de détecter et corriger certaines erreurs d'étiquetage, quand l'étiquette est incompatible avec la terminaison du mot, et de rediriger les processus de lemmatisation vers le traitement de la catégorie la plus adéquate.

Les mots traités par le lemmatiseur sont ceux dont la catégorie fait partie des catégories majeures fléchies : VCJ, SBC, ADJ, DTN. Le programme est composé de trois modules :

Conversion des catégories : Un filtre se charge de traduire les catégories affectées par l'étiqueteur de Brill en un ensemble d'étiquettes internes au programme. Ce système, basé sur une table de conversion, rend le programme relativement indépendant de l'étiqueteur et du format d'entrée.

3. Cet étiqueteur a été entraîné pour le français à l'INaLF (cf. (Lecomte & Paroubek, 1996)).

4. "e" est une terminaison ambiguë au sens où elle indique notamment le féminin singulier chez les adjectifs, le singulier chez les noms, l'impératif singulier ou la première/troisième personne du singulier du présent de l'indicatif/subjonctif...

Module de lemmatisation : Pour chaque mot à lemmatiser, la compatibilité terminaison/catégorie est vérifiée, après quoi la fonction correspondante est activée. Deux processus sont exécutés en parallèle : la génération de la base non fléchie du mot à partir de la base, éventuellement altérée, du mot fléchi, et le calcul des informations flexionnelles. Ces deux processus mettent en jeu les bases et les terminaisons des mots. Ainsi, si la séquence d'entrée est (1) *tiennent/VCJ*, la fonction découpe le mot selon le suffixe de 3^{ème} personne du pluriel *ent*. La forme neutre de la base est calculée (*tenir*) ; le calcul des informations flexionnelles résulte de l'intersection des traits de la base fléchie (qui marque soit l'indicatif présent, soit le subjonctif présent, soit l'impératif d'un verbe du 3^{ème} groupe) et du suffixe (qui est valide à tous les temps conjugués, sauf au futur et à l'impératif).

Sortie du lemmatiseur : Les résultats sont tout d'abord codés dans un format interne, puis convertis dans un format de sortie. Ainsi, la lemmatisation de (1) génère tout d'abord la séquence :

(2) VCJ tenir, 3ppPSTSUBJ/IND, (3rd group),

puis (2) est traduit en (3), où les informations flexionnelles ambiguës sont factorisées en ensembles disjoints :

(3) tiennent/VCJ:3p:pl:pst:{ind|subj}/tenir:3g

L'application du programme à (\mathcal{P}_{tag}) produit en sortie (\mathcal{P}_{lem}) :

(\mathcal{P}_{lem}) < txt id = 006232 >< ph id = 9 >La/DTN:f:s/le part/SBC:..s/part respective/ADJ:f:s/respectif des/DTC:..p/du deux/CAR/deux portions/SBC:..p/portion du/DTC:m:s/du système/SBC:..s/système racinaire/ADJ:..s/racinaire ... varie/VCJ: {{1|3}p:s:pst:{ind|subj}|2p:s:pst:imper}/varier:1g ... < /ph >< /txt >

Les performances du lemmatiseur ont été évaluées en comparant ses résultats avec ceux des deux corpus lemmatisés manuellement : on trouve 1,75% de différence par rapport au lexique TLFnome⁵ (412 081 entrées) et 0,12% de différence par rapport à un corpus d'articles de "Le Monde" étiqueté par l'Équipe TALANA (432 636 occurrences). La comparaison n'a porté que sur les phénomènes couverts actuellement par le programme. Entre autres, les taux calculés ne tiennent pas compte des noms composés par hyphénation, dont les règles de lemmatisation n'ont pas encore été intégrées au programme. Essentiellement, les autres phénomènes non encore couverts sont : le nombre de certains noms ou adjectifs en fonction du suffixe et la détermination du genre de la majorité des noms par l'examen de leur terminaison quand cela est pertinent.

3.2. Pré-traitement du thésaurus

Le thésaurus est converti en une liste de termes. Il est ensuite étiqueté par l'étiqueteur de Brill qui a été au préalable entraîné spécifiquement pour l'étiquetage de séquences nominales. Le taux de réussite est de 98%. Après la lemmatisation, la structure d'un terme est la suivante :

006467 système/système:SBC:sg racinaire/racinaire:ADJ:sg

3.3. L'indexation des textes par les termes

L'indexation des textes par les termes consiste à :

- acquérir de nouveaux termes à partir des textes lemmatisés (Étape 2 de la Figure 2) pour enrichir le thésaurus (voir § 3.3.1),
- utiliser le thésaurus ainsi enrichi pour extraire les variations des termes et indexer les textes (Étape 3 de la Figure 2, voir § 3.3.2).

5. Lexique de référence dérivé de *le Trésor de la Langue Française*

3.3.1. Acquisition de termes à partir du corpus

Les termes recyclés obtenus à partir du thésaurus AGROVOC sont de bons descripteurs mais malheureusement ils sont en nombre insuffisant pour permettre une réelle indexation du document. Un grand nombre de termes spécifiques au domaine reflétés par le corpus manquent ainsi que des termes à caractère générique, c'est-à-dire non spécifique au domaine mais porteur dans notre corpus d'informations essentielles. Ces nouveaux termes sont extraits automatiquement⁶ de notre corpus à l'aide du programme ACABIT (Daille, 1996) basé sur une approche mixte. La méthodologie est la suivante : une série de filtres linguistiques chargés de repérer les séquences de mots partageant une structure morphosyntaxique caractéristique des termes est appliquée sur le corpus étiqueté et lemmatisé. Ces structures sont ensuite soumises à un test statistique qui permettent de les classer de la plus à la moins pertinente. Ces filtres linguistiques sont exprimés sous forme de grammaires locales chargées de détecter les noms composés binaires, c'est-à-dire comportant deux mots n'appartenant pas à l'ensemble des mots fonctionnels⁷. Cette décision de se concentrer sur des termes nominaux binaires est fondée sur des critères linguistiques et statistiques, et sur un critère pratique : ils sont en adéquation avec le format d'entrée de FASTR (cf. section 3.3.2).

Les règles (2) et (3) ci-dessous exprimées à l'aide de la syntaxe des expressions régulières isolent des patrons prédéfinis de termes binaires : la règle (1) détecte des groupes adjectivaux ; la règle (2) des termes de structure morphosyntaxique SBC ADJ comme par exemple *alimentation hydrique* ; la règle (3) des termes de structure Nom1-Préposition-Nom2 comme par exemple *eau du sol*.

- | | | |
|-----|--------|----------------------|
| (1) | adjp | ((ADV* (ADJ VPAR)+)+ |
| (2) | nadj | SBC adjp |
| (3) | nprepn | (SBC nadj) PREP SBC |

où VPAR représente un participe passé et PREP une préposition.

Ce programme d'extraction de termes prend aussi en compte certaines de leurs variations. Pour illustrer la manière dont sont prises en compte les variations des termes, examinons la règle (4) permettant de reconnaître une coordination de tête entre deux termes de structure Nom1-Préposition-Nom2 :

- | | | |
|-----|---------------|---------------------|
| (4) | nprep_ncoordn | nprepn COO PREP SBC |
|-----|---------------|---------------------|

La règle (4) isole deux termes binaires, par exemple *système de surface* et *système de profondeur* au sein d'une séquence de mots comportant une coordination comme par exemple *système racinaire de surface et de profondeur*.

Tous les candidats termes binaires ainsi extraits du corpus sont ensuite, pour chaque phrase, triés selon la valeur calculée du coefficient de vraisemblance (loglike) (Dunning, 1993). Cette mesure statistique donne de bons résultats pour obtenir un classement des candidats termes, extraits à l'aide des filtres linguistiques, du plus au moins représentatif du domaine (Daille *et al.*, 1995). Pour chaque phrase ne sont retenus que deux candidats termes ; ceux pour qui la valeur du coefficient de vraisemblance est la plus élevée. Cette étape est illustrée par la table 1(a). Ce filtrage ad-hoc nous permet de couvrir correctement le corpus puisque 60 % des candidats-termes sont ainsi retenus. Néanmoins, si le coefficient de vraisemblance donne de bons résultats

6. L'extraction des termes à partir du corpus est automatique ; néanmoins la liste de candidats termes proposés nécessite habituellement d'être post-éditée de manière à éliminer les mauvais candidats.

7. Sont considérés comme mots fonctionnels, les prépositions, les articles, les conjonctions de coordinations pouvant apparaître à l'intérieur d'un terme.

Candidat terme		Score	Candidat terme	Correct	Incorrect
<i>système</i>	<i>racinaire</i>	174.819	600	575 (96 %)	25 (4%)
<i>eau</i>	<i>sol</i>	79.1731	Type d'erreur		
<i>alimentation</i>	<i>hydrique</i>	68.6754	Préposition composée		
<i>disponibilité</i>	<i>eau</i>	63.6223	Mauvais étiquetage		
<i>fonction</i>	<i>demande</i>	6.47268	Mauvais rattachement de l'adjectif		
<i>part</i>	<i>respective</i>	3.13626	Quantifieur		
<i>alimentation</i>	<i>arbre</i>	2.29382	Divers		
<i>système</i>	<i>profondeur</i>	1.70805			
<i>système</i>	<i>surface</i>	0.00165			

(a) Sortie d'Acabit sur (*P*)

(b) Précision des candidats termes retenus par phrase

TAB. 1 – *L'extraction des termes par ACABIT (S_{lem})*

pour obtenir un tri des candidats en fonction de leur représentativité par rapport à un domaine, nous n'avons pour l'instant aucune preuve qu'il soit aussi pertinent pour juger leur caractère informatif. L'évaluation de la pertinence du coefficient de vraisemblance pour l'accès à l'information est actuellement en cours et fait intervenir des documentalistes, experts dans le domaine de l'agro-alimentaire. Il est aussi évalué s'il est plus intéressant d'effectuer le filtrage au niveau de la phrase ou directement au niveau du paragraphe (ce dernier représentant une référence bibliographique).

Dans l'attente des résultats de l'évaluation sur le caractère informatif des candidats termes, nous avons ici vérifié leur correction linguistique. La précision du programme d'extraction a été évaluée sur un ensemble de 300 phrases extraites au hasard du corpus. La table 1(b) montre que 95% des candidats termes retenus sont bien formés.

3.3.2. Extraction de variantes pour l'indexation

L'étape suivante de la chaîne de traitement consiste à indexer les documents en recherchant dans les documents les occurrences de termes et de leurs variantes (voir le récapitulatif de l'organisation du projet à la figure 2). L'indexation est faite au moyen de l'analyseur transformationnel partiel FASTER (Jacquemin *et al.*, 1997).

L'analyse des documents par FASTER repose sur les trois types de données suivants :

1. un corpus morphologiquement analysé et désambiguïsé où les mots fléchis sont associés à un lemme unique et une structure de traits morphologiques unique. Cette structure attributs/valeurs définit les caractéristiques morphologiques du mot fléchi telles que le temps, le genre, le nombre, etc. (voir § 3.1);
2. une liste de termes lemmatisés où les mots ont un lemme unique et une catégorie syntaxique unique. Soit ces termes appartiennent à un thésaurus — et on parle alors d'indexation contrôlée (voir § 3.2) —, soit ces termes sont issus d'ACABIT, l'outil d'acquisition automatique de termes présenté au § 3.3.1 — et il s'agit alors d'indexation libre —;
3. une métagrammaire du langage considéré décrivant les patrons de variation terminologique (environ 100 métrarègles).

Dans ce paragraphe, nous décrivons le rôle de la métagrammaire qui implémente des transformations morpho-syntaxiques locales représentant des variations terminologiques. Les métrarègles sont des fonctions de l'ensemble des termes dans l'ensemble des variantes ; elles prennent

Variante no	Terme	Variante	Type
(2.a)	<i>système racinaire</i>	<i>système racinaire</i>	0
(2.b)	<i>système de racine</i>	<i>système racinaire</i>	NàA
(2.c)	<i>système de surface</i>	<i>système racinaire de surface</i>	Modif
(2.d)	<i>alimentation hydrique</i>	<i>alimentation hydrique</i>	0
(2.e)	<i>alimentation de arbre</i>	<i>alimentation hydrique d'un arbre</i>	Modif
(2.f)	<i>variation de alimentation</i>	<i>alimentation hydrique d'un arbre varie</i>	NàV
(2.g)	<i>disponibilité en eau</i>	<i>disponibilité de l'eau</i>	Synap
(2.h)	<i>eau de sol</i>	<i>eau du sol</i>	0

TAB. 2 – La sortie de l'analyseur FASTER sur la phrase (\mathcal{P}) du § 3

en entrée un terme de la base et produisent en sortie un patron de variation de ce terme. Ce patron est partiellement instancié puisqu'il comporte des éléments non lexicaux. Il est utilisé pour retrouver dans les documents des occurrences de variantes de termes. La table 2 illustre les index en sortie de FASTER lors de l'analyse de la phrase (\mathcal{S}) du § 3 à partir des termes acquis par ACABIT sur le corpus [AGR]. On retrouve, bien sûr, en sortie, les deux termes acquis par ACABIT sur cette phrase (les deux premières lignes de la table 1(a) et les occurrences (2.a) et (2.h) de la table 2). Les autres occurrences correspondent à des termes acquis sur d'autres phrases du corpus [AGR].

La plupart des travaux en TALN pour la recherche d'information sont appliqués en indexation libre (Schwarz, 1990; Sheridan & Smeaton, 1992; Strzalkowski, 1996). Ces analyseurs à large couverture décomposent des structures syntaxiques en dépendances élémentaires qui constituent les index du texte. Au contraire, la finalité de FASTER est l'indexation contrôlée : il s'agit de retrouver, au moyen d'une base de termes et d'une métagrammaire de variations locales, les occurrences de ces termes et de leurs variantes. La sortie de l'analyseur sont des liens linguistiquement documentés aux termes de la base. Une étiquette 0 sur le lien signifie qu'il s'agit d'une occurrence — éventuellement fléchie — d'un terme de la base. Toutes les autres étiquettes dénotent des variantes.

Les exemples de la table 2 illustrent les deux familles de variations actuellement prises en compte par FASTER :

- Les *variantes syntaxiques* mettent en jeu une transformation de la syntaxe du terme sans modifications morphologiques des mots autres que les flexions. Ainsi, la variante (2.c) de la table 2, est obtenue au moyen d'une variation, appelée modification, qui est l'insertion d'un modifieur adjectival après le nom tête dans un terme de structure initiale Nom-Préposition-Nom. Cette variation permet d'extraire *système racinaire de surface* comme une variante de *système de surface*.
- Les *variantes morpho-syntaxiques* contiennent au moins un mot du terme initial qui a subi une transformation de morphologie dérivationnelle. En outre, ces variations peuvent également faire appel à des transformations syntaxiques. Par exemple, la variante (2.f) de la table 2 est obtenue au moyen d'une transformation Nom à Verbe. Celle-ci permet d'extraire *alimentation hydrique d'un arbre varie* comme une variante de *variation de alimentation*.

L'extraction de la variante (2.f) repose sur une transformation morpho-syntaxique implémentée par la métarègle suivante :

$$\begin{aligned}
 & \text{Métarègle NàV(SBC1 PREP2 SBC3)} & (1) \\
 & = \text{SBC3 (ADJ? (PREP DTN? ADJ? SBC (ADJ (COO ADJ)?)?)} \\
 & \quad (\text{COO DTN? ADV? ADJ? SBC ADJ?})? \text{Vaux? Vaux? ADV?} \text{ V1 :} \\
 & < \text{SBC1 root ref} > = < \text{V1 root ref} > .
 \end{aligned}$$

Variante no	Terme	Variante
(3.a)	détermination de teneur	teneur en chrome est également déterminée
(3.b)	utilisation de chromatographie	chromatographie liquide est utilisée
(3.c)	traitement de plante	plantes traitées
(3.d)	analyse de tige	tiges ont été analysés
(3.e)	élévation de teneur	teneur élevée
(3.f)	utilisation de le chambre	chambres de refroidissement utilisées
(3.g)	développement de sol	sols développés
(3.h)	influence de variété	variété est influencée
(3.i)	apport de azote	azote apporté
(3.j)	détermination de facteur	facteurs déterminent

TAB. 3 – Les dix premières variantes Nom à Verbe de [AGR]

Termes	Variantes syntaxiques			Variantes morpho-syntaxiques					
	Coor	Modif	Synap	NàV	NàN	NàA	AàN	AàV	AàAv
55812 (60.1%)	2165	9983	4216	7259	9213	2663	1414	90	2
	16364 (17.6%)			20641 (22.3%)					
	92817								

TAB. 4 – Évaluation quantitative de l'extraction terminologique sur [AGR]

où COO est une conjonction de coordination, ADV un adverbe et Vaux un verbe auxiliaire.

La métarègle (1) stipule qu'un terme de structure Nom1-Préposition-Nom2 peut donner lieu à une réalisation verbale sous les trois conditions suivantes :

1. la tête du syntagme prépositionnel (ici *alimentation*) est le sujet du syntagme verbal,
2. la tête du terme (ici *variation*) est morphologiquement liée à la tête du syntagme verbal (ici *varie*),
3. la structure syntaxique du syntagme verbal et son sujet doivent appartenir aux chaînes décrites par l'expression régulière (2) suivante :

$$\begin{aligned} & \text{SBC (ADJ? (PREPDTN? ADJ? SBC (ADJ (COO ADJ)?)?)?} & (2) \\ & \text{(COO DTN? ADV? ADJ? SBC ADJ)?Vaux? Vaux? ADV?) V.} \end{aligned}$$

La métarègle (1) permet d'extraire 3 100 variantes du corpus [AGR] à partir de la liste de termes extraits par ACABIT. La table 3 illustre les 10 premières variantes ainsi extraites.

La quantité et la qualité de l'extraction terminologique effectuées par FASTER sont évaluées, du point de vue linguistique, dans les tables 4 et 5. La précision est mesurée sur un échantillon de 1 000 variations choisies aléatoirement. La table 4 indique que les variantes de termes représentent 40% des index. Elles sont ainsi réparties : 45% des variantes sont des variantes syntaxiques et les 55% restant sont des variantes morpho-syntaxiques. Alors que les occurrences de termes (non variantes) sont extraites avec une très grande précision, les variantes de termes sont obtenues avec une précision d'environ 75%.

3.4. Classification: SDOC et la méthode des mots associés

Le dernier élément de la chaîne exploite les termes obtenus au § 3.3 pour construire des classes de termes. Ces classes visent à synthétiser le contenu informatif du corpus [AGR] et sont construites suivant la méthode des mots associés, une approche robuste utilisée dans les systèmes de recherche d'information (voir (Callon *et al.*, 1986; Callon *et al.*, 1991; Salton *et al.*, 1994)). Cette méthode, basée sur la distribution dans les textes des unités d'information que sont

Termes	Variantes syntaxiques			Variantes morfo-syntaxiques					
	Coord	Modif	Synap	NàV	NàN	NàA	AàN	AàV	AàAv
611/611 (100%)	19/24	81/104	24/27	45/63	84/121	21/28	15/18	0/1	2/2
	124/155 (80%)			167/233 (72%)					
	291/388 (75%)								

TAB. 5 – Précision de l'extraction terminologique sur [AGR] (taux d'index corrects)

les termes est mise en œuvre dans le système SDOC (Grivel & Francois, 1995; Grivel *et al.*, 1995).

La classification des termes prend en compte deux critères : leur fréquence dans le corpus et la cooccurrence des termes dans chacun des résumés. SDOC procède en deux étapes :

1. construction d'un réseau d'association basé sur la cooccurrence des termes en calculant un indice de similarité appelé Indice d'Équivalence : $E_{ij} = C_{ij}^2 / (C_i \times C_j)$, où i et j sont deux termes, C_i and C_j leur fréquence, et C_{ij} la fréquence de leur cooccurrence.
2. partition du réseau en utilisant l'algorithme du simple lien (voir (Callon *et al.*, 1986)).

Les experts du domaine ont montré que les classes qui constituent le réseau lexical sont des structures sémiotiquement décodables, comme le montre la classe *racine* de la Figure 1 (voir aussi (Muller *et al.*, 1997)). Ces classes mettent en valeur des relations de type hypo/hyponymie (*évapotranspiration/évaporation*) ; « partie-de » (*racine/pommier*) ; complémentarité (*racine/partie aérienne*) ; synonymie (*racine/système racinaire*) ainsi que des relations processus/entité (*enracinement/système racinaire*).

4. Perspectives

Le système présenté ici est une association originale et unique de techniques de traitement automatique de la langue et d'infométrie. Elle a permis la réalisation d'un outil d'extraction, de représentation et d'accès aux concepts d'un domaine extraits d'un très grand volume de textes. Ce projet est actuellement opérationnel sur un corpus français de résumés dans le domaine de l'agriculture. La prise en compte d'une nouvelle langue (anglais) et de nouveaux domaines techniques seront terminés dans quelques mois.

Le système est actuellement utilisé par des experts pour l'analyse de l'information en agriculture et en médecine. De plus, une évaluation de la pertinence des candidats termes extraits lors des étapes (3) et (4) pour l'indexation est en cours de réalisation par des indexeurs.

La réalisation de ce système nous a également permis de mettre au point un lemmatiseur pour le français, langage considéré traditionnellement comme trop complexe pour permettre une lemmatisation qui n'est pas basée sur l'utilisation d'un dictionnaire.

Les applications que nous envisageons pour un tel système sont la recherche d'information, l'enrichissement automatique des thésaurus, et l'assistance pour l'indexation manuelle et la catégorisation de documents.

Références

- BESCHERELLE (1990). *La Conjugaison 12000 Verbes*. Paris: Hatier.
- BRILL E. (1993). *A Corpus-Based Approach to Language Learning*. PhD thesis, University of Pennsylvania.

- CALLON M., COURTIAL J.-P. & LAVILLE F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research : the case of polymer chemistry. *Scientometrics*, **22**(1), 155–205.
- M. CALLON, J. LAW & A. RIP, Eds. (1986). *Mapping the Dynamics of Science and Technology*. London: Macmillan Press.
- DAILLE B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In J. L. KLAVANS & P. RESNICK, Eds., *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, p. 49–66. MIT Press.
- DAILLE B., GAUSSIER E. & LANGÉ J.-M. (1995). An evaluation of statistical scores for word association. The Tbilisi Symposium on Language, Logic and Computation. Tbilisi, Géorgie.
- DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1).
- GRIVEL L. & FRANCOIS C. (1995). *Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique*, p. 81–113. Jean-Max Noyer, presses universitaires de rennes edition.
- GRIVEL L., MUTSCHKE P. & POLANCO X. (1995). Thematic mapping on bibliographic databases by cluster analysis : a description of the sdoc environment with solis. *Journal of Knowledge Organization*, **22**(2), 70–77.
- 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. In *Actes de TALN'97, Grenoble, France*.
- JACQUEMIN C., KLAVANS J. L. & TZOUKERMANN E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings, 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL'97), Madrid*, p. 24–31: ACL.
- KARTTUNEN L. (1994). Constructing lexical transducers. In *15ème conférence internationale Coling 94, Kyoto, Japan*, p. 406–411. Vol. I.
- LECOMTE J. & PAROUBEK P. (1996). *Le catégoriseur d'Eric Brill. Mise en œuvre de la version entraînée à l'INaLF*. Rapport interne, CNRS–INaLF, Nancy.
- MULLER C., POLANCO X., ROYAUTÉ J. & TOUSSAINT Y. (1997). *Acquisition et structuration de connaissances en corpus : éléments méthodologiques*. Rapport interne RR-3198, Rapport de Recherche INRIA. 45 pages, format postscript accessible : “ftp.inria.fr”.
- SALTON G., ALLAN L. & BUCKLEY C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, **37**(2), 97–108.
- SAVOY J. (1993). Stemming of french words based on grammatical categories. *JASIS: Journal of the American Society for Information Sciences*, **44**(1), 1–9.
- SCHWARZ C. (1990). Automatic syntactic analysis of free text. *Journal of the American Society for Information Science*, **41**(6), 408–417.
- SHERIDAN P. & SMEATON A. F. (1992). The application of morpho-syntactic language processing to effective phrase matching. *Information Processing & Management*, **28**(3), 349–369.
- STRZALKOWSKI T. (1996). Natural language information retrieval. *Information Processing & Management*, **31**(3), 397–417.