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

Yannick Toussaint[1], Fiammetta Namer[2],
Béatrice Daille[3], Christian Jacquemin[4],
Jean Royauté[5] et Nabil Hathout[6].

[1]LORIA-INRIA UMR 7503, Nancy, [2]Universit´e de Nancy II,
[3]IRIN, Nantes, [4]LIMSI, Orsay, [5]INIST-CNRS, Nancy, [6]INaLF-CNRS, Nancy.

Résumé

Cet article pr´esente une chaîne de traitement automatique r´ealis´ee dans le cadre du projet ILIAD (Informatique Linguistique et Infom´etrie pour l'Analyse de grands fonds Documentaires) du GIS Sciences de la Cognition. Cette chaîne est d´edi´ee à l'analyse de l'information à partir de corpus de textes de très grand volume, en français. Elle est exp´eriment´ee sur un corpus de 2,5 Mb et a conduit à la cr´eation de 50 classes de termes. Ces classes sont construites sur la base de la cooccurrence des termes et repr´esentent des connaissances du domaine. Les diff´erentes ´etapes de la chaîne associent des m´ethodes linguistiques informatiques et des m´ethodes statistiques : pr´e-traitement des textes, ´etiquetage, morphologie, terminologie et analyse des documents. Pour chacune d'entre elles, nous pr´esentons les m´ethodes, les outils ainsi que leur ´evaluation.

1. Introduction

Cet article pr´esente 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 m´ethodes informatiques linguistiques et des m´ethodes statistiques, assurant, pour chacune des ´etapes, une grande robustesse. De plus, son architecture ouverte facilite la prise en compte de nouvelles fonctionnalit´es par l'insertion de nouveaux modules de traitement.

L'analyse de l'information peut être d´efinie comme un ensemble d'outils et de m´ethodes permettant à un op´erateur humain de collecter l'information contenue dans un corpus sans le lire de façon s´equentielle. Cet op´erateur peut être ainsi capable d'analyser l'avancement d'un domaine scientifique ou technique en consid´erant simplement un ensemble de classes de termes, construites à partir du corpus initial.

ILIAD a ´et´e exp´eriment´e sur un corpus de 2,5 Mb de textes en français sur le domaine de l'agriculture. Ce sont des r´esum´es de notices bibliographiques et constituent de ce fait, un corpus d'une grande homogen´eït´e. L'analyse par ILIAD produit 50 classes actuellement utilis´ees par les experts pour la veille technologique ou la recherche d'information.

Après une pr´esentation de l'architecture globale du système, nous d´ecrirons les diff´erentes ´etapes du traitement, les m´ethodologies adopt´ees (bas´ees sur les avanc´ees r´ecentes de la termi-

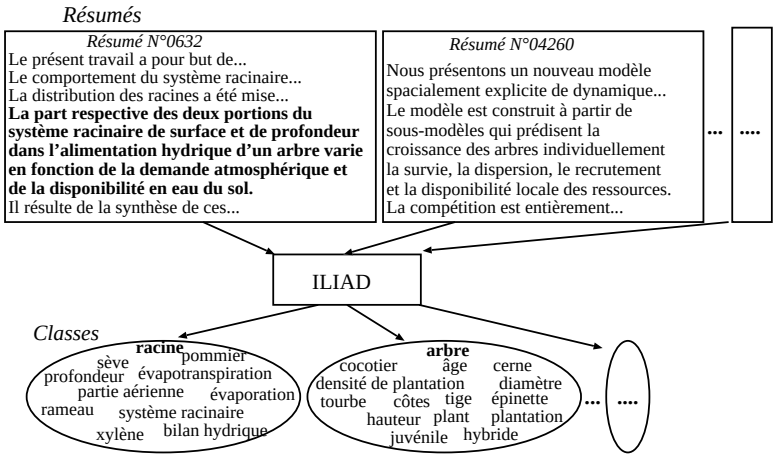

FIG. 1 – *Du texte à la classification*

nologie, tant du point de vue linguistique que du point de vue de la repr´esentation des connaissances) puis leur ´evaluation.

2. Architecture

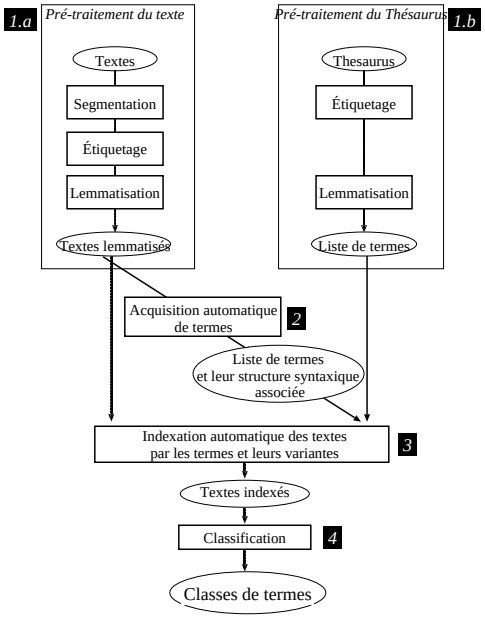

FIG. 2 – *Architecture globale*

La Figure 1 donne un exemple de classes construites à partir d'un ensemble de r´esum´es. Chaque classe regroupe, sur la base d'une mesure de similarit´e, les termes fr´equemment cooccurrents dans les textes. Ce processus est divis´e en quatre ´etapes (voir Figure 2): (1.a) et (1.b) pour le pr´e-traitement des textes et du th´esaurus (´etiquetage morpho-syntaxique, segmentation et lemmatisation), puis (2) pour l'acquisition des termes, (3) pour l'indexation, et (4) pour la classification.

Chaque ´etape repose sur des outils ou m´ethodes existants (ACABIT, FASTER, SDOC) qui sont d´ecrits dans les sections suivantes. ILIAD a ´et´e appliqu´e à un corpus de 7272 r´esum´es de textes en français (2,5 Mb) sur le domaine de l'agriculture, extraits de la base PASCAL [1] (corpus d´esign´e par [AGR]), et utilise le th´esaurus AGROVOC[2] qui est une taxinomie d'environ 15 000 termes, associ´es à leurs synonymes et enregistr´es dans un format proche de la norme SGML.

1. PASCAL est la base documentaire scientifique d´evelopp´ee et maintenue par l'INIST-CNRS, France
2. Th´esaurus multilingue d´evelopp´e par AGRIS - FAO (Unit´e 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 th´esaurus sont pr´epar´es par deux processus identiques, c'est pourquoi nous ne d´etaillons que celui concernant les textes.

Ceux-ci sont tout d'abord segment´es en phrases, et mis dans un format proche de SGML, puis l'´etiqueteur de Brill[3](1993) associe à chaque mot sa cat´egorie 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 cat´egorie des d´eterminants, SBC des noms, DTC des pr´epositions contract´ees, ADJ des adjectifs, CAR des cardinaux, et VCJ des verbes conjugu´es.

La dernière ´etape dans la pr´eparation du texte est la lemmatisation des mots fl´echis, au moyen d'un programme bas´e sur des règles, qui associe à chaque forme fl´echie 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 interpr´etation d´epend de la cat´egorie du mot auquel elles sont attach´ees.

De nombreux lemmatiseurs du français, utilis´es en recherche d'information sont, pour cette raison, bas´es sur la consultation d'un dictionnaire (cf. (Savoy, 1993)). D'autres systèmes, bas´es, 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-´etiquet´e. Par cons´equent, le programme d'ILIAD, qui prend en entr´ee le texte ´etiquet´e, ne doit pas se pr´eoccuper de d´esambiguïsation. Ne d´ependant pas d'un dictionnaire (sauf pour quelques listes d'exceptions aux règles utilis´ees), il est capable de prendre en compte les mots inconnus et les n´eologismes qui sont suppos´es avoir un comportement flexionnel r´egulier. Enfin, de par sa conception, le lemmatiseur est capable de d´etecter et corriger certaines erreurs d'´etiquetage, quand l'´etiquette est incompatible avec la terminaison du mot, et de rediriger les processus de lemmatisation vers le traitement de la cat´egorie la plus ad´equate.

Les mots trait´es par le lemmatiseur sont ceux dont la cat´egorie fait partie des cat´egories majeures fl´echies : VCJ, SBC, ADJ, DTN. Le programme est compos´e de trois modules :

Conversion des catégories : Un filtre se charge de traduire les cat´egories affect´ees par l'´etiqueteur de Brill en un ensemble d'´etiquettes internes au programme. Ce système, bas´e sur une table de conversion, rend le programme relativement ind´ependant de l'´etiqueteur et du format d'entr´ee.

3. Cet ´etiqueteur a ´et´e entraˆın´e pour le franç ais `a l'INaLF (cf. (Lecomte & Paroubek, 1996)).

4. "e" est une terminaison ambigu¨e au sens o`u elle indique notamment le f´eminin singulier chez les adjectifs, le singulier chez les noms, l'imp´eratif singulier ou la premi`ere/troisi`eme personne du singulier du pr´esent de l'indicatif/subjonctif…

Module de lemmatisation : Pour chaque mot à lemmatiser, la compatibilit´e terminaison/cat´egorie est v´erifi´ee, après quoi la fonction correspondante est activ´ee. Deux processus sont ex´ecut´es en parallèle : la g´en´eration de la base non fl´echie du mot à partir de la base, ´eventuellement alt´er´ee, du mot fl´echi, et le calcul des informations flexionnelles. Ces deux processus mettent en jeu les bases et les terminaisons des mots. Ainsi, si la s´equence d'entr´ee est (1) *tiennent/VCJ*, la fonction d´ecoupe le mot selon le suffixe de 3ème personne du pluriel *ent*. La forme neutre de la base est calcul´ee (*tenir*) ; le calcul des informations flexionnelles r´esulte de l'intersection des traits de la base fl´echie (qui marque soit l'indicatif pr´esent, soit le subjonctif pr´esent, soit l'imp´eratif d'un verbe du 3ème groupe) et du suffixe (qui est valide à tous les temps conjugu´es, sauf au futur et à l'imp´eratif).

Sortie du lemmatiseur : Les r´esultats sont tout d'abord cod´es dans un format interne, puis convertis dans un format de sortie. Ainsi, la lemmatisation de (1) g´enère tout d'abord la s´equence :

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

puis (2) est traduit en (3), où les informations flexionnelles ambiguës sont factoris´ees 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 ´et´e ´evalu´ees en comparant ses r´esultats avec ceux des deux corpus lemmatis´es manuellement : on trouve 1,75% de diff´erence par rapport au lexique TLFnome[5] (412 081 entr´ees) et 0,12% de diff´erence par rapport à un corpus d'articles de "Le Monde" ´etiquet´e par l'Équipe TALANA (432 636 occurrences). La comparaison n'a port´e que sur les ph´enomènes couverts actuellement par le programme. Entres autres, les taux calcul´es ne tiennent pas compte des noms compos´es par hyph´enation, dont les règles de lemmatisation n'ont pas encore ´et´e int´egr´ees au programme. Essentiellement, les autres ph´enomènes non encore couverts sont : le nombre de certains noms ou adjectifs en fonction du suffixe et la d´etermination du genre de la majorit´e des noms par l'examen de leur terminaison quand cela est pertinent.

3.2. Pré-traitement du thésaurus

Le th´esaurus est converti en une liste de termes. Il est ensuite ´etiquet´e par l'´etiqueteur de Brill qui a ´et´e au pr´ealable entrain´e sp´ecifiquement pour l'´etiquetage de s´equences nominales. Le taux de r´eussite 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 :

a) acqu´erir de nouveaux termes à partir des textes lemmatis´es (Étape 2 de la Figure 2) pour enrichir le th´esaurus (voir § 3.3.1),

b) utiliser le th´esaurus 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 r´ef´erence d´eriv´e de *le Trésor de la Langue Française*

3.3.1. Acquisition de termes á partir du corpus

Les termes recycl´es obtenus à partir du th´esaurus AGROVOC sont de bons descripteurs mais malheureusement ils sont en nombre insuffisant pour permettre une r´eelle indexation du document. Un grand nombre de termes sp´ecifiques au domaine refl´et´es par le corpus manquent ainsi que des termes à caractère g´en´erique, c'est-à-dire non sp´ecifique au domaine mais porteur dans notre corpus d'informations essentielles. Ces nouveaux termes sont extraits automatiquement [6] de notre corpus à l'aide du programme ACABIT (Daille, 1996) bas´e sur une approche mixte. La m´ethodologie est la suivante : une s´erie de filtres linguistiques charg´es de rep´erer les s´equences de mots partageant une structure morphosyntaxique caract´eristique des termes est appliqu´ee sur le corpus ´etiquet´e et lemmatis´e. 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 exprim´es sont forme de grammaires locales charg´ees de d´etecter les noms compos´es binaires, c'est-à-dire comportant deux mots n'appartenant pas ´a l'ensemble des mots fonctionnels[7]. Cette d´ecision de se concentrer sur des termes nominaux binaires est fond´e sur des critères linguistiques et statistiques, et sur un critère pratique : ils sont en ad´equation avec le format d'entr´ee de FASTR (cf. section 3.3.2).

Les règles (2) et (3) ci-dessous exprim´ees à l'aide de la syntaxe des expressions r´egulières isolent des patrons pr´ed´efinis de termes binaires : la règle (1) d´etecte 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-Pr´eposition-Nom2 comme par exemple *eau du sol* .

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

où VPAR repr´esente un participe pass´e et PREP une pr´eposition.

Ce programme d'extraction de termes prend aussi en compte certaines de leurs variations. Pour illustrer la manière dont sont prises en comptes les variations des termes, examinons la règle (4) permettant de reconnaître une coordination de tête entre deux termes de structure Nom1-Pr´eposition-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 s´equence 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, tri´es selon la valeur calcul´ee du coefficient de vraisemblance (loglike) (Dunning, 1993). Cette mesure statistique donne de bons r´esultats pour obtenir un classement des candidats termes, extraits à l'aide des filtres linguistiques, du plus au moins repr´esentatif 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 ´elev´ee. Cette ´etape est illustr´ee par la table 1(a). Ce filtrage ad-hoc nous permet de couvrir correctement le corpus puisque 60 % des candidats-termes sont ainsi retenus. N´eanmoins, si le coefficient de vraisemblance donne de bons r´esultats

6. L'extraction des termes `a partir du corpus est automatique ; n´eanmoins la liste de candidats termes propos´es n´ecessite habituellement d'ˆetre post-´edit´ee de mani`ere `a ´eliminer les mauvais candidats.

7. Sont consid´er´es comme mots fonctionnels, les pr´epositions, les articles, les conjonctions de coordinations pouvant apparaˆıtre `a l'int´erieur d'un terme.

Candidat terme		Score
système	***racinaire***	**174.819**
eau	***sol***	**79.1731**
alimentation	*hydrique*	68.6754
disponibilité	*eau*	63.6223
fonction	*demande*	6.47268
part	*respective*	3.13626
alimentation	*arbre*	2.29382
syst`eme	*profondeur*	1.70805
syst`eme	*surface*	0.00165

(a) Sortie d'Acabit sur (P)

Candidat terme	Correct	Incorrect
600	575 (96 %)	25 (4%)

Type d'erreur	Nombre
Pr´eposition compos´ee	12 (48 %)
Mauvais ´etiquetage	7 (28 %)
Mauvais rattachement de l'adjectif	3 (12 %)
Quantifi eur	1 (4 %)
Divers	2 (8 %)

(b) Pr´ecision des candidats termes retenus par phrase

TAB. 1 – *L'extraction des termes par ACABIT ($\mathcal{S}_{lem}$)*

pour obtenir un tri des candidats en fonction de leur repr´esentativit´e par rapport à un domaine, nous n'avons pour l'instant aucune preuve qu'il soit aussi pertinent pour juger leur caractère informatif. L'´evaluation 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 ´evalu´e s'il est plus int´eressant d'effectuer le filtrage au niveau de la phrase ou directement au niveau du paragraphe (ce dernier repr´esentant une r´ef´erence bibliographique).

Dans l'attente des r´esultats de l'´evalation sur le caratère informatif des candidats termes, nous avons ici verifi´e leur correction linguistique La pr´ecision du programme d'extraction a ´et´e ´evalu´ee sur un ensemble de 300 phrases extraites au hasard du corpus. La table 1(b) montre que 95% des candidats termes retenus sont bien form´es.

3.3.2. Extraction de variantes pour l'indexation

L'´etape 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 r´ecapitulatif 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 donn´ees suivants :

1. un corpus morphologiquement analys´e et d´esambiguïs´e où les mots fl´echis sont associ´es à un lemme unique et une structure de traits morphologiques unique. Cette structure attributs/valeurs d´efinit les caract´eristiques morphologiques du mot fl´echi telles que le temps, le genre, le nombre, etc. (voir § 3.1);

2. une liste de termes lemmatis´es où les mots ont un lemme unique et une cat´egorie syntaxique unique. Soit ces termes appartiennent à un th´esaurus — et on parle alors d'indexation contrôl´ee (voir § 3.2) —, soit ces termes sont issus d'ACABIT, l'outil d'acquisition automatique de termes pr´esent´e au § 3.3.1 — et il s'agit alors d'indexation libre —;

3. une m´etagrammaire du langage consid´er´e d´ecrivant les patrons de variation terminologique (environ 100 m´etarègles).

Dans ce paragraphe, nous d´ecrivons le rôle de la m´etagrammaire qui impl´emente des transformations morpho-syntaxiques locales repr´esentant des variations terminologiques. Les m´etarègles sont des fonctions de l'ensemble des termes dans l'ensemble des variantes ; elles prennent

Variante no	Terme	Variante	Type
(2.a)	*syst`eme racinaire*	*syst`eme racinaire*	0
(2.b)	*syst`eme de racine*	*syst`eme racinaire*	N`aA
(2.c)	*syst`eme de surface*	*syst`eme racinaire de surface*	Modif
(2.d)	*alimentation hydrique*	*alimentation hydrique*	0
(2.e)	*alimentation de arbre*	*alimentation hydrique d'un arbre*	Modif
(2.f)	*variation de alimentation*	*alimentation hydrique d'un arbre varie*	N`aV
(2.g)	*disponibilité en eau*	*disponibilité de l'eau*	Synap
(2.h)	*eau de sol*	*eau du sol*	0

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

en entr´ee un terme de la base et produisent en sortie un patron de variation de ce terme. Ce patron est partiellement instanci´e puisqu'il comporte des ´el´ements non lexicaux. Il est utilis´e 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 appliqu´es en indexation libre (Schwarz, 1990; Sheridan & Smeaton, 1992; Strzalkowski, 1996). Ces analyseurs à large couverture d´ecomposent des structures syntaxiques en d´ependances ´el´ementaires qui constituent les index du texte. Au contraire, la finalit´e de FASTER est l'indexation contrôl´ee : il s'agit de retrouver, au moyen d'une base de termes et d'une m´etagrammaire de variations locales, les occurrences de ces termes et de leurs variantes. La sortie de l'analyseur sont des liens linguistiquement document´es aux termes de la base. Une ´etiquette 0 sur le lien signifie qu'il s'agit d'une occurrence — ´eventuellement fl´echie — d'un terme de la base. Toutes les autres ´etiquettes d´enotent 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, appel´ee modification, qui est l'insertion d'un modifieur adjectival après le nom tête dans un terme de structure initiale Nom-Pr´eposition-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 d´erivationnelle. En outre, ces variations peuvent ´egalement 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 impl´ement´ee par la m´etarègle suivante :

```
Métarègle NàV( SBC1 PREP2 SBC3 )                                    (1)
    = SBC3 (ADJ? (PREP DTN? ADJ? SBC (ADJ (COO ADJ)?)?)?
         (COO DTN? ADV? ADJ? SBC ADJ?)? Vaux? Vaux? ADV?) V1 :
< SBC1 root ref > = < V1 root ref >.
```

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	N`aV	N`aN	N`aA	A`aN	A`aV	A`aAv
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 m´etarègle (1) stipule qu'un terme de structure Nom1-Pr´eposition-Nom2 peut donner lieu à une r´ealisation verbale sous les trois conditions suivantes :

1. la tête du syntagme pr´epositionnel (ici *alimentation*) est le sujet du syntagme verbal,
2. la tête du terme (ici *variation*) est morphologiquement li´ee à la tête du syntagme verbal (ici *varie*),
3. la structure syntaxique du syntagme verbal et son sujet doivent appartenir aux chaînes d´ecrites par l'expression r´egulière (2) suivante :

SBC (ADJ? (PREPDTN? ADJ? SBC (ADJ (COO ADJ)?)?)? (2)
(COO DTN? ADV? ADJ? SBC ADJ?)?Vaux? Vaux? ADV?) V.

La m´etarègle (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 quantit´e et la qualit´e de l'extraction terminologique effectu´ees par FASTER sont ´evalu´ees, du point de vue linguistique, dans les tables 4 et 5. La pr´ecision est mesur´ee sur un ´echantillon de 1 000 variations choisies al´eatoirement. La table 4 indique que les variantes de termes repr´esentent 40% des index. Elles sont ainsi r´eparties : 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 pr´ecision, les variantes de termes sont obtenues avec une pr´ecision d'environ 75%.

3.4. Classification: SDOC et la méthode des mots associés

Le dernier ´el´ement de la chaîne exploite les termes obtenus au § 3.3 pour construire des classes de termes. Ces classes visent à synth´etiser le contenu informatif du corpus [AGR] et sont construites suivant la m´ethode des mots associ´es, une approche robuste utilis´ee dans les systèmes de recherche d'information (voir (Callon *et al.*, 1986; Callon *et al.*, 1991; Salton *et al.*, 1994)). Cette m´ethode, bas´ee sur la distribution dans les textes des unit´es d'information que sont

Termes	Variantes syntaxiques			Variantes morpho-syntaxiques					
	Coor	Modif	Synap	N`aV	N`aN	N`aA	A`aN	A`aV	A`aAv
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 fr´equence dans le corpus et la cooccurrence des termes dans chacun des r´esum´es. SDOC procède en deux ´etapes :

1. construction d'un r´eseau d'association bas´e sur la cooccurrence des termes en calculant un indice de similarit´e appel´e 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 fr´equence, et C_{ij} la fr´equence de leur cooccurrence.

2. partition du r´eseau en utilisant l'algorithme du simple lien (voir (Callon *et al.*, 1986)).

Les experts du domaine ont montr´e que les classes qui constituent le r´eseau lexical sont des structures s´emiotiquement d´ecodables, 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/hyperonymie (*évapotranspiration/évaporation*) ; « partie-de » (*racine/pommier*) ; compl´ementarit´e (*racine/partie aérienne*); synonymie (*racine/système racinaire*) ainsi que des relations processus/entit´e (*enracinement/système racinaire*).

4. Perspectives

Le système pr´esent´e ici est une association originale et unique de techniques de traitement automatique de la langue et d'infom´etrie. Elle a permis la r´ealisation d'un outil d'extraction, de repr´esentation et d'accès aux concepts d'un domaine extraits d'un très grand volume de textes. Ce projet est actuellement op´erationnel sur un corpus français de r´esum´es dans le domaine de l'agriculture. La prise en compte d'une nouvelle langue (anglais) et de nouveaux domaines techniques seront termin´es dans quelques mois.

Le système est actuellement utilis´e par des experts pour l'analyse de l'information en agriculture et en m´edecine. De plus, une ´evaluation de la pertinence des candidats termes extraits lors des ´etapes (3) et (4) pour l'indexation est en cours de r´ealisation par des indexeurs.

La r´ealisation de ce système nous a ´egalement permis de mettre au point un lemmatiseur pour le français, langage consid´er´e traditionnellement comme trop complexe pour permettre une lemmatisation qui n'est pas bas´ee sur l'utilisation d'un dictionnaire.

Les applications que nous envisageons pour un tel système sont la recherche d'information, l'enrichissement automatique des th´esaurus, et l'assistance pour l'indexation manuelle et la cat´egorisation 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. Tbilissi, G´eorgie.

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 franç ais encore plus robuste en traitant fi nement les mots inconnus avec les donn´ees 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 *15`eme conf´erence internationale Coling 94, Kyoto,Japan*, p. 406–411. Vol. I.

LECOMTE J. & PAROUBEK P. (1996). *Le cat´egoriseur d'Eric Brill. Mise en œuvre de la version entraîn´ee `a l'INaLF*. Rapport interne, CNRS–INaLF, Nancy.

MULLER C., POLANCO X., ROYAUTÉ J. & TOUSSAINT Y. (1997). *Acquisition et structuration de connaissances en corpus : ´el´ements m´ethodologiques*. 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 fi les. *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.