Etude comparative de deux outils d'acquisition de termes complexes

Agata Savary

LADL, Université de Marne-la-Vallée
5, bd Descartes, 77454 Marne-la-Vallée Cedex 2
xsavary@free.fr

Résumé

Nous comparons deux outils d'extraction de termes complexes : l'un fondé sur une analyse symbolique accompagnée d'un calcul statistique, et l'autre sur l'analyse symbolique et l'utilisation de ressources terminologiques initiales. Les résultats d'extraction obtenus par les deux outils à partir d'un corpus de 280 000 mots révèlent entre autres l'importance quantitative des termes hapax et des termes contenant plus de deux constituants, ainsi que l'intérêt de l'adaptation de ressources terminologiques existantes au traitement automatique.

1. Introduction

Nous avons effectué une étude comparative de deux méthodes d'acquisition de termes complexes à partir d'un corpus spécialisé: ACABIT (DAILLE 1994), et LexProTerm (CHROBOT 1999) basé sur les fonctionnalités du système INTEX (SILBERZTEIN 1999).

Nous présentons les méthodes formelles employées par ces deux outils, et nous comparons les résultats d'acquisition obtenus sur un grand corpus du domaine de l'informatique.

2. Motivation et état de l'art

Comme le démontre LEHRBERGER (1986), un langage spécialisé (*sublanguage*) n'est pas un sous-ensemble du langage naturel « standard ». L'intersection des deux est d'habitude non vide, mais il existe des phrases qui appartiennent au langage standard et non pas au langage spécialisé et inversement. En conséquence, le traitement automatique du langage spécialisé devrait être effectué par des outils adaptés au domaine du texte traité. Mais une telle approche, considérée comme trop coûteuse, est rarement admise dans les applications TALN existantes. La tendance est plutôt contraire : proposer des systèmes applicables à chaque domaine technique sans nécessité d'adaptation. Si cette solution est « robuste », elle ne devrait servir, selon nous, que comme outil intermédiaire pour l'élaboration d'autres outils adaptés au domaine. Par exemple, un extracteur terminologique « général » peut servir à l'élaboration ou l'enrichissement de lexiques et grammaires spécialisés.

L'un des outils permettant de rendre compte de la spécificité du domaine traité dans le cadre du traitement automatique est un lexique électronique spécialisé. Il peut atteindre un niveau de précision de description aussi élevé que chez GROSS et GUENTHNER (1998). Un lexique spécialisé contenant des informations uniquement morphologiques est employé dans la deuxième méthode d'acquisition de termes que nous analysons, LexProTerm. La première méthode présentée, ACABIT, est indépendante du domaine.

Le problème de recherche de termes dans des textes a une bibliographie très riche grâce aux enjeux qu'elle représente : l'indexation automatique des textes, la recherche documentaire, la création de lexiques et dictionnaires uni- et multilingues, la traduction automatique et l'aide à la traduction, etc. Les travaux existants dans ce domaine peuvent être classés selon différents critères, comme ceux admis par JACQUEMIN (1997, pp. 8-31). Pour le but de notre analyse il est utile de mentionner au moins les trois critères suivants :

1) La reconnaissance des termes (ou indexation) contre l'acquisition (ou extraction) des termes.

   Le premier type d'outil sert à retrouver dans des textes les termes déjà connus, i.e. existant dans une liste contrôlée. Pour cela, il suffit de disposer d'un outil d'analyse lexicale tenant compte des unités complexes, comme INTEX (SILBERZTEIN 1993), qui est à la base de LexProTerm. La difficulté de réalisation plus fine de cette tâche provient du fait qu'un terme peut apparaître sous différentes variantes (e.g. *birth date = date of birth*). La variation terminologique a été traitée par JACQUEMIN (1997) et JACQUEMIN, KLAVANS et TZOUKERMANN (1997). Du côté de l'extraction des termes nous pouvons classer un nombre important de travaux, comme le système Termino de DAVID et PLANTE (1990), Xtract de SMADJA (1993), ANA de ENGUEHARD et PANTERA (1994), JUSTESON et KATZ (1995), Lexter de BOURIGAULT (1994), FASTER de JACQUEMIN (1997 ; la deuxième partie de son étude), BOWDEN et al. (1998), ainsi que les deux systèmes faisant l'objet de cette étude.

2) L'extraction statistique des termes contre l'extraction par analyse symbolique.

   La première approche est fondée sur l'idée que les mots qui forment une unité lexicale ont tendance à apparaître ensemble plus souvent que d'autres combinaisons de mots. Les travaux de référence sont par exemple ceux de ENGUEHARD et PANTERA (1994), JUSTESON et KATZ (1995), NAKAGAWA et MORI (1998). La deuxième approche, visant le repérage de toutes les occurrences de termes, non seulement les plus fréquentes, est représentée par DAVID et PLANTE (1990), BOURIGAULT (1994), JACQUEMIN (1997), LADOUCER et COCHRANE (1996), ainsi que LexProTerm. Des modèles hybrides qui emploient conjointement des outils linguistiques (étiqueteur lexicaux, grammaires locales, analyseurs syntaxiques) et statistiques, sont par exemple celui de SMADJA (1993), et ACABIT.

3) L'extraction « initiale » de termes contre l'enrichissement terminologique.

   La plupart des outils d'extraction adoptent cette première approche, i.e. ils admettent le corpus et éventuellement un outil linguistique général (étiqueteur grammatical, grammaire locale, ou analyseur syntaxique) comme les seuls points de départ pour la recherche de termes. La deuxième approche tient compte de l'existence d'une base terminologique initiale comme point de départ pour la recherche de nouveaux termes. Elle est représentée par JACQUEMIN (1997), ainsi que partiellement par ENGUEHARD et PANTERA (1994), et est aussi admise dans LexProTerm.

3. Comparaison des méthodes

Selon les critères présentés plus haut nous allons analyser :

- d'une part une méthode d'extraction initiale, fondée sur une analyse symbolique accompagnée d'un calcul statistique (ACABIT),
- d'autre part une méthode d'enrichissement terminologique qui, par le moyen de l'analyse symbolique, effectue à la fois la reconnaissance de termes contrôlés et l'acquisition de nouveaux termes (LexProTerm).

Le tableau Tab. 1 présente la comparaison des principes de fonctionnement des deux algorithmes.

<table>
<tr><th>Phases de l'extraction</th><th>ACABIT</th><th>LexProTerm</th></tr>
<tr><td>Ressources à l'entrée du programme</td><td>Corpus spécialisé.
Dictionnaire de mots composés généraux (taille non indiquée).
Etiqueteur basé sur 720 règles probabilistes concernant des suffixes de mots simples et sur une liste de 7200 exceptions.</td><td>Corpus spécialisé.
Dictionnaire général 166 000 lemmes simples et 60 000 lemmes composés.
Dictionnaire spécialisé (pour le domaine de l'informatique : 73 000 lemmes simples et 58 000 lemmes composés).</td></tr>
<tr><td></td><td colspan="2">Patrons syntaxiques de syntagmes nominaux sous forme d'automates finis.</td></tr>
<tr><td>Etiquetage du texte</td><td>Chaque mot du texte obtient à l'issue d'une analyse statistique une seule étiquette grammaticale.</td><td>Chaque mot (sauf les mots grammaticaux) obtient toutes les étiquettes grammaticales qui lui correspondent dans les lexiques utilisés.</td></tr>
<tr><td>Patrons syntaxiques</td><td>Représentent des syntagmes nominaux généraux.</td><td>Représentent des syntagmes nominaux constitués, pour la plupart, de termes simples et complexes déjà connus, ou de mots non existants dans les dictionnaires utilisés (supposés être des néologismes du domaine).</td></tr>
<tr><td>Recherche de candidats termes</td><td colspan="2">Les patrons syntaxiques sont appliqués au texte.</td></tr>
<tr><td></td><td>Différentes variantes du même candidat terme sont rattachées.
Seuls sont retenus les candidats binaires (pouvant éventuellement apparaître dans le corpus à l'intérieur de syntagmes plus longs) avec la fréquence supérieure à 1.
Le candidats sont ensuite triés selon la valeur statistique appelée coefficient de vraisemblance.</td><td>Toutes les séquences extraites sont retenues.
Elles sont présentées dans l'ordre de leur fréquence d'apparition dans le texte.</td></tr>
</table>

Tab. 1. Comparaison des deux d'algorithmes d'acquisition de termes

4. Comparaison des résultats

4.1. Difficulté d'évaluation

Toute évaluation d'un outil d'extraction doit être précédée de la réflexion sur ce qui doit être considéré comme un terme pertinent. Comme le proposent BOURIGAULT et SLODZIAN (1999), il convient d'abandonner l'idée que la terminologie d'un domaine est un

ensemble de notions prédéfinies et correspondant aux concepts rigides qu'il suffit de découvrir. La terminologie doit plutôt être *construite* individuellement pour chaque nouvelle application, et en ce sens c'est une terminologie textuelle, et non pas métalinguistique. Une telle *contruction* de terminologie d'un texte doit être effectuée par un terminologue accompagné d'un expert du domaine. Dans l'expérience que nous décrivons ci-dessous nous avons joué ces deux rôles à la fois. C'est pourquoi les résultats que nous présentons doivent être admis avec précaution.

4.2. Précision

Afin de pouvoir comparer les performances des deux outils d'acquisition terminologique, nous avons choisi un corpus (IBM 1997-99) portant sur l'architecture des ordinateurs. Il contient 280 000 formes simples, soit plus de 400 pages (1,68 mégaoctets) de texte.

Le tableau Tab. 2 contient un résumé des résultats obtenus par les deux algorithmes. Nous nous servons du critère de *précision* définie comme la proportion de bons termes parmi tous les candidats termes proposés par un extracteur. Quant au taux de *rappel* (la proportion de bons termes proposés par un extracteur parmi tous les termes existant dans le texte traité), nous n'avons pas pu effectuer son calcul pour un si grand corpus (pour ceci nous aurions dû d'abord marquer manuellement toutes les occurrences de termes sur 400 pages de texte). Néanmoins, nous pouvons obtenir certains indices sur le *silence* (les termes pertinents existant dans le texte et non extraits) propre aux deux méthodes si nous analysons les candidats pertinents qui ont été extraits par l'un de ces outils et non pas par l'autre (voir la section 4.5).

Extracteur	**Candidats extraits**	**Candidats pertinents**	**Précision**
ACABIT	2 133	1 425	67%
LexProTerm	22 760	10 677 (13% connus)	47%
LexProTerm (fréquence = 1)	18 216	7 601 (10% connus)	42%
LexProTerm (fréquence > 1)	4 551	3 104 (20% connus)	68%
Candidats communs	1 432	1 163	81%

Tab. 2. Précision des deux outils

4.3. Résultats d'ACABIT

ACABIT a trouvé 2 133 candidats-termes différents dans le corpus de test. Nous les avons analysés manuellement et nous en avons retenu 1 423 comme termes valables, d'où le taux de précision égal à 67%.

Parmi les candidats pertinents extraits la grande majorité est constituée de termes contenant deux unités lexicales pleines (noms, adjectifs, adverbes, verbes). En effet ACABIT se concentre sur l'extraction de ce type de termes appelés « termes de base », et son grand avantage est de relier différentes variantes des mêmes termes. Par exemple, les séquences suivantes :

[1] *permanent failure, permanent failures, permanent physical failure, permanent single-bit failure, failure is permanent*

donnent lieu à l'extraction un seul terme : *permanent failure*. Ce résultat est basé entre autres sur l'hypothèse qu'un terme de longueur supérieure à 2 est souvent obtenu à partir des termes de longueur 1 ou 2 par une des deux opérations, selon la terminologie admise par DAILLE (1994) : l'insertion (insertion de modifieurs ou substitution) ou la juxtaposition (surcomposition, placement de modifieurs en position non initiale). Par exemple, *permanent physical failure* peut être obtenu à partir de *permanent failure* soit par insertion de *physical* :

[2] *permanent failure + physical ⇒ permanent physical failure*

soit par la substitution de *failure* par *physical failure* :

[3] *permanent failure* (*failure ← physical failure*) ⇒ *permanent physical failure*

ACABIT permet aussi d'extraire certains candidats termes qui ont plus de deux unités lexicales pleines car une séquence de mots reliés par un tiret est automatiquement considérée comme une seule unité. Parmi les candidats pertinents extraits par ACABIT nous retrouvons 203 termes de ce type :

[4] *test-case generator, Cache-to-cache latency, data-pattern-dependent jitter*

En analysant les candidats non pertinents extraits par ACABIT, nous retrouvons :

- des syntagmes libres, tels que (le symbole « ° » signifie que la séquence qui le suit n'est pas un candidat-terme pertinent, indépendamment du fait qu'elle est ou non un syntagme correct de l'anglais):

  [5] *°great deal, °correct result, °brief description*

- des syntagmes non nominaux :

  [6] *°be braodcast, °at the end, °available through IBM, °integrated cryptographic*

- des séquences placées à la frontière de deux syntagmes (le candidat extrait est souligné) – ceci est probablement dû aux erreurs de l'étiquetage :

  [7] *…tester-based °diagnostics work best when the failure is in the scan chain…*

- un grand nombre de séquences binaires qui sont des sous-séquences de termes valables plus larges (le candidat extrait est souligné). Ceci est dû à deux phénomènes. Premièrement, un terme composé de longueur supérieure à 2 n'est pas forcément toujours obtenu par une insertion ou une juxtaposition de termes binaires. Par exemple, les termes suivants :

  [8] *most significant bit, direct attached crypto*

  sont obtenus, respectivement, par un placement d'un adjectif non terminologique *most significant* devant un terme unaire *bit*, et par une élision du terme simple *operations* dans le terme quaternaire *direct attached crypto operations*. Dans les deux termes ternaires ci-dessus ACABIT extrait donc, à tort, °*significant bit* et °*attached crypto* comme candidats-termes de base.

  Deuxièmement, un terme peut être une insertion ou une juxtaposition de termes binaires, mais ACABIT n'a pas correctement déterminé quels composants constituent le terme binaire d'origine. Par exemple, dans la surcomposition suivante :

  [9] *operation-graphe + finite-state machine ⇒ operation-graphe finite-state machine*

  le candidat °*operation-graphe machine* a été extrait à tort.

Nous avons pu constater, pour certains termes valables de longueur 2 extraits par ACABIT, que l'intérêt de leur repérage était mineur par rapport aux termes plus longs qui les contenaient. Par exemple, les deux candidats suivants :

[10] °*integrated cluster*, *cluster bus*

ont été extraits, l'un à tort et l'autre à raison. Toutes les 16 occurrences pour chacune de ces deux séquences ont lieu à l'intérieur du même terme ternaire :

[11] *integrated cluster bus*

dont le fort statut terminologique est confirmé par l'existence de l'abréviation *ICB*. Puisque aucune de ces sous-séquences binaires n'apparaît indépendamment l'une de l'autre, nous croyons qu'il serait plus convenable de considérer le terme ternaire comme terme de base, et non pas comme une insertion ou juxtaposition de termes binaires. De la même façon devraient être considérés comme termes de base les séquences suivantes :

[12] *Integrated Cryptographic Facility* (*ICRF*), *random number generator* (*RNG*), *absolute address history table* (*AAHT*), *key agreement protocol* (*KEP*), etc.

et non pas leurs sous-séquences, extraites par ACABIT, qui n'apparaissent qu'à l'intérieur des ces premières :

[13] °*Cryptographic Facility*, °*number generator*, °*address history*, °*history table*, °*agreement protocol*, etc.

4.4. Résultats de LexProTerm

Nous avons analysé manuellement la liste de candidats-termes extraits par LexProTerm et la liste de leurs occurrences dans le corpus de test IBM (1997-99). Le nombre de candidats était de 22 766, et le nombre total d'occurrences de 35 118. Un candidat était classé comme pertinent s'il apparaissait au moins une fois dans le corpus en tant que terme valable.

Plus de 22 000 candidats à analyser manuellement constitue une tâche assez lourde dans le procès d'extraction de termes. Il est possible qu'un utilisateur choisisse de ne tenir compte que des candidats les plus fréquents. Si l'on fixe le seuil de fréquence à 2, comme ceci est le cas dans ACABIT, il ne reste que 4 551 candidats à valider et le taux de précision atteint 68%. Mais le prix à payer est une baisse considérable de rappel, car les termes apparaissant une seule fois sont plus de deux fois plus nombreux (7601) que les autres (3104).

Environs 44% de candidats pertinents extraits par LexProTerm contiennent plus de deux constituants, et la longueur maximale obtenue est de 7 constituants :

[14] *architectural-level instruction stream test-case generator*

L'analyse des candidats non pertinents extraits par LexProTerm permet d'observer plusieurs phénomènes étant à l'origine du bruit (candidats extraits à tort) et du silence :

– Certaines occurrences de termes connus ne sont pas pertinentes, par exemple le terme *new ligne* signifie un caractère à un code particulier, mais dans le contexte ci-dessus cette séquence apparaît en tant que syntagme libre :

  [15] *…in order to bring the new line into the BCE and operand buffers…*

– La qualité de ressources lexicographiques utilisées a une grande influence sur la qualité des résultats de l'extraction. Les dictionnaires spécialisés utilisés par LexProTerm contenaient quelques mots simples de la langue générale qui n'ont pas de sens particulier dans le domaine de l'informatique, comme :

[16] *aspect, problem, issue, typical, appropriate, important,* etc.

Ces mots provoquent soit l'extraction des syntagmes libres (°*important aspects,* °*performance issue*) et des mots composés généraux (°*person-month*), soit le marquage incorrect de frontières des termes (°*appropriate ABIST macros*, °*host adapter portion*) où seules les séquences soulignées devraient être extraites.

– Par le choix de méthode d'extraction (recherche de syntagmes nominaux constitués de termes déjà connus) LexProTerm n'est pas à l'abri des cas ou un mot commun fréquent en anglais a un sens spécialisé dans le domaine technique traité. Par exemple, le nom *key* apparaît en position de modifieur dans 30 syntagmes nominaux extraits à tort, et dans 48 syntagmes extraits à raison. Dans ce premier cas, par exemple dans :

[17] °*key decision,* °*key internal cache management concepts*

key a son sens commun - « le plus important, le principal » - indépendamment du sens du nom qu'il modifie. Dans le deuxième cas il concerne le concept d'une clef dans le domaine de la cryptographie :

[18] *key generation, key translation, key attribute, key enabler*

– Le manque de désambiguïsation de mots est à l'origine d'un grand ensemble de candidats non pertinents. Souvent la frontière d'un terme est mal repérée à cause d'un verbe à la troisième personne du singulier ambigu avec un nom spécialisé au pluriel, comme dans l'exemples ci-dessous :

[19] *The control flow includes command/status buses and finite-state machines.*

Finalement, il est important d'admettre que nous avons eu de nombreux doutes quant au statut de certains candidats, surtout ceux contenant des mots simples qui semblent intermédiaires entre la langue générale et la langue spécialisée, comme :

[20] *mechanism, coverage, element, performance, failure, component*

Leurs occurrences dans les séquences extraites semblaient affaiblir sensiblement la pertinence de ces séquences (alors que les sous-séquences soulignées sont sûrement des termes) :

[21] *bus-snooping mechanism, dc stuck-at-fault test coverage, coupled-system performance, host hardware component*

4.5. Comparaison

Le tableau Tab. 2 permet de comparer la précision des résultats d'ACABIT et de LexProTerm. Remarquons, que LexProTerm, après l'introduction du seuil de fréquence fixé à deux, arrive à extraire deux fois plus de termes qu'ACABIT, avec la même précision que ce dernier. Il est important de remarquer que seulement 20% des candidats pertinents de LexProTerm sont des termes reconnus grâce à leur présence dans les dictionnaires utilisés.

Il y a 262 termes extraits par ACABIT et non pas par LexProTerm. Près de la moitié de ce nombre sont des termes binaires qui apparaissent dans des termes plus larges extraits par LexProTerm. Il reste 135 termes que LexProTerm n'a pas extraits pour diverses raisons :

– ils apparaissent dans des candidats non pertinents extraits par LexProTerm, comme *emulation code* (LexProTerm a extrait °*control unit emulation code take* et °*emulation code reduces*), *guard-band range* (dans °*specified guard-band ranges*),

– ils contiennent des mots qui ne sont pas des termes connus; c'est le cas de *multifiber*

ferrule, *jitter budget*, *decoupling capacitance*, *correctable error*, etc.

– leur structure syntaxique n'est pas prévue dans le patron de recherche de LexProTerm, comme dans le cas de *mode of operation*, *fencing command* (les structures prépositionnelles et les participes présents des verbes ne sont pas admis dans le patron, sauf s'ils apparaissent dans des termes controlés).

Les termes extraits par LexProTerm et non pas par ACABIT sont très nombreux (environs 9 500) et il est difficile de les classer tous. Nous en mentionnons seulement quelques types les plus fréquents. Compte tenu des principes de fonctionnement d'ACABIT, il est normal que l'ensemble de termes extraits par LexProTerm et non pas par ACABIT contienne :

– presque tous les termes de longueur supérieure à deux,

– presque tous les termes qui apparaissent une seule fois dans le corpus.

ACABIT		**LexProTerm**	
+	bonne précision	+	bon rappel
+	indépendant du domaine (i.e. n'exige pas de liste initiale de termes)	-	exige une liste initiale de termes du domaine, les résultats de l'extraction dépendent de la qualité de cette liste
-	exige un grand corpus	+	indépendant de la taille du corpus
-	n'effectue que l'extraction initiale	+	permet de repérer les termes connus d'une façon sûre et d'enrichir leur liste
+	bons résultats pour les termes binaires	-	beaucoup de bruit au niveau des candidats binaires
-	mauvais résultats pour les termes avec plus de 2 constituants	+	extrait les termes maximaux
+	extrait des termes avec une préposition	-	extrait peu de termes avec une préposition
+	lemmatise les constituants des candidats termes et rattache différentes formes fléchies du même terme	-	différentes formes fléchies du même terme sont considérées comme termes indépendants
+	relie les termes avec leurs variantes	-	seules les séquences contiguës sont recherchées dans le texte ; différentes variantes sont considérées comme termes séparés

Tab. 3. Comparaison des deux outils d'extraction

Quant aux termes binaires avec la fréquence supérieure à 1, LexProTerm en a extrait plus de mille de plus qu'ACABIT (remarquons qu'un terme n'a pas toujours la même fréquence pour LexProTerm que pour ACABIT, notamment dans les cas où un terme binaire peut faire partie d'un terme plus long, ou bien quand il apparaît sous différentes variantes). Nous y trouvons par exemple *array macro*, *access key*, *active configuration*, *address conflict*, etc. Ces termes n'ont probablement pas été retenus par ACABIT à cause de leurs faibles valeurs du coefficient de vraisemblance.

En résumé de ce test comparatif, le tableau Tab. 3 rassemble les avantages (marqués «+») et les inconvénients (marqués «-») des deux méthodes d'extraction.

5. Conclusions

Le test comparatif d'ACABIT et LexProTerm semble avoir démontré que l'utilisation de ressources terminologiques initiales peut considérablement améliorer certains aspects de l'acquisition de termes. Il a également confirmé l'hypothèse, admise dans ce dernier outil, que la création d'un nouveau terme se fait souvent par une combinaison grammaticalement correcte de termes simples et composés déjà existants. Les résultats obtenus ont aussi soulevé deux problèmes :

– celui des termes hapax (à fréquence 1) dont le nombre dans des textes est deux fois plus élevé que d'autres termes ; ceci implique que les méthodes statistiques d'extraction, bonnes au niveau de la précision, ne peuvent pas être satisfaisantes au niveau du rappel,

– celui des termes contenant plus que deux constituants qui sont presque aussi nombreux dans des textes que les termes binaires, contrairement à ce que constate B. DAILLE (1996, p. 127), et qu'il faut donc traiter au même titre que ces derniers.

LexProTerm permet de remédier à ces deux problèmes et d'obtenir des résultats riches d'extraction s'il dispose d'un dictionnaire couvrant précisément le domaine traité, et si la terminologie disponible dans ce dictionnaire est assez complète et de bonne qualité. Néanmoins, il demande aussi des améliorations, telles que l'introduction d'un étiqueteur grammatical. Afin qu'un outil de ce type puisse s'appliquer à un domaine technique, il a besoin de ressources terminologiques spécialisées. Ces ressources existent pour de nombreux domaines – ce sont les dictionnaires techniques traditionnels (i.e. destinés à un lecteur humain) qu'il faut convertir en des formats utilisables par des programmes informatiques. Même si cette conversion n'est pas entièrement automatisable (SAVARY 2000, pp. 90-92, 114-117), il serait important d'étudier les possibilités de son automatisation partielle.

Remerciements

Je voudrais remercier Béatrice Daille et Chantal Enguehard pour leur aide dans la réalisation de cette étude.

Références

BOURIGAULT D. (1994), *LEXTER un Logiciel d'Extraction de Terminologie. Application à l'extraction des connaissances à partir de textes.* Thèse de doctorat en Mathématiques, Informatique Appliquée aux Sciences de l'Homme, Paris, École des Hautes Études en Sciences Sociales.

BOURIGAULT D., SLODZIAN M . (1999). Pour une terminologie textuelle. In Ch. ENGUEHARD, A. CONDAMINES. Eds., *Terminologies Nouvelles, N° 19, Actes du Colloque « Terminologie et Intelligence Artificielle TIA-99 », Nantes 10-11 mai 1999,* décembre 1998 - juin 1999, Bruxelles, Agence de la francophonie et Communauté française de Belgique.

BOWDEN P., LINDSAY E., HALSTEAD P. (1998), Automatic Acronym Acquisition in a Knowledge Extraction Program. In *Proceedings from COMPUTERM, the First Workshop on Computational Terminology, August 15, 1988,* University of Montreal.

CHROBOT A. (1999). Enrichissement terminologique en anglais fondé sur des dictionnaires généraux et spécialisés. In Ch. ENGUEHARD, A. CONDAMINES. Eds., *Terminologies Nouvelles, N° 19, Actes du*

Colloque « Terminologie et Intelligence Artificielle TIA-99 », Nantes 10-11 mai 1999, décembre 1998 - juin 1999, Bruxelles, Agence de la francophonie et Communauté française de Belgique.

DAILLE B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*. Thèse de doctorat en Informatique Fondamentale, Université Paris 7.

DAILLE B. (1996). ACABIT : une maquette d'aide à la construction automatique de banques terminologiques. In A. CLAS, P. THOIRON, BEJOINT Eds., *Lexicomatique et Dictionnairique*, FMA, Beyrouth, pp. 123-136.

DAVID S., PLANTE P. (1990). De la nécessité d'une approche morpho-syntaxique en analyse de textes. In *ICO* 2(3), pp. 140-155, Québec.

ENGUEHARD Ch., PANTERA L. (1994). Automatic Natural Acquisition of Terminology. In *Journal of Quantitative Linguistics*, Vol. 2, No. 1, pp. 27-32, Netherland, Swets & Zeitlinger.

GROSS G., GUENTHNER F. (1998). Traitement automatique des domaines. In *Revue française de linguistique appliquée*, III-2, pp. 47-56.

IBM (1997-99). IBM S/390 Server G3/G4, dans *IBM Journal of Research and Development*, Vol.41, No. 4/5, 1997 et IBM S/390 Server G5/G6, dans *IBM Journal of Research and Development*, Vol.43, No. 5/6, 1999, *www.research.ibm.com/journal*

JACQUEMIN Ch. (1997). *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Habilitation à diriger des recherches en informatique, IRIN, Université de Nantes.

JACQUEMIN Ch., KLAVANS J., TZOUKERMANN E. (1997). Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Barcelona, 7-10 July 1997*, Association for Computational Linguistics.

JUSTESON J., KATZ S. (1995). Technical terminology : some linguistic properties and an algorythm for identification in text. In *Natural Language Engineering, 1(1)*, pp. 9-27.

LADOUCEUR J., COCHRANE G., (1996). Termplus, système d'extraction terminologique. In M. GRARSON, Ed., *Terminologies nouvelles, Banques de terminologie, Actes de la table ronde, Québec, 18 et 19 janvier 1996*, N°15, Bruxelles, pp. 52-56.

LEHRBERGER J. (1986). Sublanguage Analysis. In R. GRISHMAN, R. KITTREDGE Eds., *Analyzing Language in Restricted Domains*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

NAKAGAWA H., MORI T. (1998). Nested Collocation and Compound Noun For Term Extraction. In *Preceedings from COMPUTERM, the First Workshop on Computational Terminology, August 15, 1988*, University of Montreal.

SAVARY A. (2000). *Recensement et description des mots composés - méthodes et applications*. Thèse de doctorat en Informatique Fondamentale, Université de Marne-la-Vallée.

SILBERZTEIN M. (1999). Intex: a FST toolbox. In *Theoretical Computer Science*, Elsevier Sciences.

SMADJA F. (1993). XTRACT: An Overview. In *Computers and the Humanities*, 26, pp. 399-413, Kluwer Academic Publishers, Netherlands.