Sélection de termes dans un dictionnaire d'informatique : Comparaison de corpus et critères lexico-sémantiques

Marie-Claude L'Homme

Observatoire de linguistique Sens-Texte (OLST)
Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec)
H3C 3J7
CANADA
marie-claude.l'homme@umontreal.ca

Résumé
Le présent article propose une méthode de sélection des termes devant faire partie d'un dictionnaire spécialisé et, plus précisément, un dictionnaire fondamental d'informatique. La méthode repose principalement sur un ensemble de critères lexico-sémantiques appliqués à un corpus spécialisé. Elle tient également compte de la fréquence et de la répartition des unités dans ce corpus. Dans ce travail, nous avons voulu savoir jusqu'à quel point des techniques de comparaison de corpus permettaient de ramener des termes coïncidant avec la liste obtenue par l'application des critères. L'examen de la liste générée automatiquement montre qu'un peu plus de 50 % des unités classées comme étant spécifiques par la métrique sont également retenues par le terminographe. Les résultats révèlent que la technique revêt un intérêt dans la mesure où elle permet d'aligner des choix sur des données extraites de corpus. Toutefois, la sélection automatique recèle un certain nombre d'imperfections qui doivent être corrigées par une analyse terminographique.

1. Introduction

La sélection des termes devant faire partie d'un dictionnaire spécialisé pose toujours problème et est rarement abordée de front par les terminographes. Il suffit pour s'en convaincre de comparer les contenus des dictionnaires portant sur le même domaine, par exemple l'informatique : si certains termes (ex. *ordinateur, mémoire, bit*) se retrouvent invariablement dans tous les dictionnaires, d'autres recevront un traitement nettement moins uniforme (ex. *exécuter, central*, *aide*).

Le présent article propose une méthode de sélection des entrées dans un dictionnaire fondamental d'informatique. Les objectifs du projet ainsi que le contenu envisagé du dictionnaire sont donnés à la section 2. La méthode repose principalement sur un ensemble de critères lexico-sémantiques appliqués aux unités d'un corpus spécialisé, mais elle tient également compte de la fréquence et de la répartition des termes dans ce corpus. Ces critères sont décrits à la section 3.

Nous avons voulu savoir si des techniques automatiques pouvaient venir en aide au terminographe pour mener à bien cette sélection. Nous nous sommes penchée plus précisément sur des méthodes de comparaison de corpus s'appuyant sur une métrique censée mesurer la notion de « spécificité lexicale ». La méthode d'extraction est abordée à la section 4. La liste de termes générée automatiquement est examinée afin de voir si les unités qu'elle contient sont des termes que le terminographe retiendrait au moyen des critères lexico-sémantiques. Les résultats de cette évaluation sont présentés à la section 5. Cette évaluation tient compte d'un corpus et de termes d'informatique français, mais elle peut sans conteste s'appliquer à d'autres domaines de spécialité et à d'autres langues.

2. Objectifs et contenu provisoire d'un dictionnaire fondamental d'informatique

Le dictionnaire envisagé s'inscrit dans ce qui semble devenir une nouvelle tradition en lexicographie spécialisée où on cherche à fournir pour chaque entrée un grand nombre d'informations de nature lexico-sémantique (Binon *et al.* 2000 ; Dancette et Réthoré 2000 ; Descamps 1976 ; Frawley 1988 ; Jousse et Bouveret 2003). Il semble que les dictionnaires élaborés de manière classique ne répondent pas à toutes les exigences de certains de leurs utilisateurs, notamment les traducteurs et les rédacteurs. En mettant l'accent sur les renseignements de nature définitoire et encyclopédique, les dictionnaires classiques ont tendance à occulter les données sur le fonctionnement linguistique des termes.

Ce projet de dictionnaire comporte quatre objectifs principaux :

a) Rendre compte des termes fondamentaux dans le domaine de l'informatique. Les *termes* sont les unités lexicales dont le sens peut être associé à un domaine spécialisé préalablement délimité. Les unités lexicales non informatiques seront écartées même si elles sont récurrentes dans les textes.
b) La liste des entrées retenues et les définitions devront refléter le plus possible le réseau lexical du domaine de l'informatique. Par exemple, les verbes et les adjectifs se combinant de façon privilégiée avec un terme de nature nominale qui a été retenu devront faire l'objet d'une entrée. De même, les actants sémantiques de termes prédicatifs retenus devront eux aussi être décrits. Enfin, les unités entrant dans une relation paradigmatique ou syntagmatique avec un terme préalablement sélectionné devront également faire l'objet d'une description.
c) Accorder une attention particulière aux multiples acceptions informatiques que peuvent avoir certaines formes lexicales. Les sens sont distingués à partir d'un certain nombre de tests contextuels et en fonction des relations sémantiques que les termes entretiennent avec d'autres termes (dérivation différentielle, cooccurrence compatible et différentielle, etc., Mel'čuk et al. 1995).
d) Décrire la structure actancielle des termes au moyen d'un système de classes regroupant les entités du domaine de l'informatique (L'Homme 2003).
e) Décrire, pour chacun des termes, l'ensemble des autres termes avec lesquels ils entretiennent une relation lexico-sémantique de nature paradigmatique ou

syntagmatique. La liste des relations sémantiques s'inspirent largement des fonctions lexicales du Dictionnaire explicatif et combinatoire, DEC (Mel'čuk et al. 1984-1999), mais l'explication que nous en donnons est vulgarisée et s'aligne sur les explications données dans la version informatisée du DEC, à savoir le DiCo (Polguère 2003). Le tableau 1 présente quelques exemples de relations retenues jusqu'à maintenant.

Mot clé	**Terme relié**	**Description de la relation sémantique Le terme 2 par rapport au terme 1**
comprimer	*compresser*	Synonymie
ordinateur	*clone*	Quasi-synonymie : plus spécifique
ordinateur	*appareil, machine*	Quasi-synonymie : plus général
tableur	*logiciel*	Hyperonymie
compatible	*incompatible*	Antonymie
archiver	*archivage1a*	Nom de même sens
hypertexte algorithme	*hypertextuel algorithmique*	Adjectif de même sens
convivial	*convivialité*	Nom de même sens
programmer1	*programmeur*	Agent
abonnement	*abonné*	Patient
numériser	*numériseur*	Instrument
charger1	*mémoire*	Lieu
afficher	*affichage1a*	Résultat
compiler	*recompiler*	De nouveau
configuration	*autoconfiguration*	De manière automatique
utilisateur	*mono-utilisateur*	Unique
logiciel	*tourner*	~ fonctionne
logiciel	*exécuter*	Qqn fait fonctionner ~
logiciel	*quitter*	Qqn cesse d'utiliser ~

Tableau 1 : Exemples de relations sémantiques retenues

3. Critères de sélection des termes

La collecte des termes et des renseignements permettant de procéder à leur description se fait à partir d'un corpus composé de 53 textes français portant sur l'informatique. Il s'agit d'un des corpus spécialisés élaborés à l'Observatoire de linguistique Sens-Texte (OLST). Les thèmes abordés dans les textes sont : l'initiation à la micro-informatique, l'Internet, le logiciel et le matériel, la programmation, les réseaux et les systèmes d'exploitation. Le nombre total de mots s'élève à environ 600 000. Le tableau 2 présente les détails du corpus.

Subdivisions du corpus d'informatique	Taille des corpus	
	Nombre de textes	Nombre de mots
Initiation à la micro-informatique	8	116 821
Internet	12	102 972
Logiciel	4	78 412
Matériel	5	41 816
Programmation et réseaux	11	38 909
Systèmes d'exploitation	13	221 104
Total	**53**	**600 034**

Tableau 2 : Corpus d'informatique

La sélection des termes repose sur un ensemble de critères lexico-sémantiques qui sont décrits ci-dessous (certains de ces critères sont décrits en détail dans L'Homme 1998 et 2003) :

a. L'unité extraite doit désigner une entité (« matériel », « logiciel », « entité de représentation », « unité de mesure » ou « animé ») du domaine de l'informatique (ex. *archive, carte, compilateur, programme, internaute, bit, ressource*)) ;
b. S'il s'agit d'unités prédicatives – verbes, nominalisations, adjectifs, etc. –, elles sont extraites si les actants renvoient à des entités du critère a. (ex. *charger* : *l'utilisateur charge un logiciel en mémoire; chaîne : chaîne de caractères*). Toutefois, la même unité prédicative peut se combiner avec des actants non spécialisés; si elle revêt le même sens avec ces autres actants, elle est éliminée (ex. le verbe *comporter* se combine avec des termes et avec des unités de sens non spécialisé, mais il conserve toujours le même sens).
c. S'il s'agit de dérivés morphologiques, ils doivent être sémantiquement apparentés à un terme sélectionné en fonction des critères a. ou b. (ex. *programme : programmer, programmable, reprogrammer, programmation*, etc. ; *archiver : archivage, archive*, etc.).
d. S'il s'agit d'une unité entrant dans une relation paradigmatique (autre qu'une relation morphologique déjà identifiée en c.) avec un terme sélectionné en fonction des critères a., b. ou c. (ex. *couper, coller, copier*), elle est extraite.

Outre ces quatre premiers critères lexico-sémantiques, nous faisons intervenir la fréquence et la répartition. Une unité lexicale peut prétendre au statut de terme si elle est utilisée dans un nombre élevé de textes liés à un domaine de spécialité.

3. Constitution d'une liste préliminaire de termes

Une liste de termes potentiels a été produite au moyen de méthodes de comparaison de corpus qui retiennent l'attention depuis quelque temps en terminologie (Ahmad et al.

1994 ; Chung 2003). Ces méthodes ont été utilisées dans d'autres applications, par exemple, la recherche d'information, la lexicographie, l'étude de genres textuels (Kilgariff 2001; Rayson and Garside 2000, entre autres).

Les méthodes de comparaison de corpus revêtent un intérêt dans le cadre du présent projet en ce sens qu'elles permettent : 1) de fonder le choix des entrées sur un corpus spécialisé (et non sur les connaissances préalables qu'un terminographe ou un spécialiste peut avoir d'un domaine) ; 2) de vérifier le critère de fréquence et de répartition évoqué à la section précédente. De plus, elles conçues pour dégager les termes simples contrairement à de nombreuses stratégies proposées au cours des dernières années qui focalisent sur les termes complexes.

Les métriques auxquelles nous avons fait appel s'inspire de travaux visant à dégager le vocabulaire spécifique d'un corpus (Lafon 1980 ; Lebart et Salem 1994) dont l'application à la terminologie a été proposée par Drouin (2003). La comparaison se fait sur des listes d'unités étiquetées et lemmatisées[1] au moyen d'un logiciel appelé *TermoStat* mis au point par Drouin (2003). Le programme distribue les unités dans trois catégories distinctes :

- les spécificités positives (SP+) : celles dont la fréquence est plus élevée que celle observée dans le corpus de référence (nous tenons pour acquis que les termes se situent dans cette catégorie);
- les formes banales (SP0) : celles dont la fréquence est la même que celle observée dans le corpus de référence;
- les spécificités négatives (SP-) : celles dont la fréquence est moins élevée que celle observée dans le corpus de référence.

Deux méthodes ont permis de dégager les spécificités lexicales des textes d'informatique :

- La première méthode consistait à comparer le corpus d'informatique au corpus journalistique (*Le Monde* 2001 qui comprend environ 30 millions de mots). Les 25 premières spécificités positives obtenues par cette méthode sont reproduites dans le tableau 3 (le tableau présente l'unité lexicale, l'étiquette morphosyntaxique attribuée par l'étiqueteur, la fréquence brute de l'unité dans le corpus d'informatique et la valeur-test attribuée par *TermoStat*).

Vingt-cinq premières SP+			
Unité	**Partie du discours**	**Fréquence brute**	**Valeur test**
fichier	SBC	3956	360.825
commande	SBC	1902	201.749
option	SBC	1486	182.338
serveur	SBC	1166	180.477
utilisateur	SBC	1117	167.307
configuration	SBC	845	162.82
utiliser	VB	1996	161.788
répertoire	SBC	1003	153.668
système	SBC	2699	152.979

disquette	SBC	609	148.431
ordinateur	SBC	1283	140.484
touche	SBC	855	138.648
logiciel	SBC	1166	137.933
imprimante	SBC	537	137.692
disque	SBC	1093	129.889
mémoire	SBC	1240	125.77
windows	SBP	613	125.066
clavier	SBC	579	122.24
caractère	SBC	1096	118.081
recommander	ADJPAR	516	117.004
linux	SBP	442	116.261
bit	SBC	382	110.72
paramètre	SBC	455	109.856
interface	SBC	412	109.64
permettre	VB	2625	108,873

Tableau 3 : Liste des 25 premières spécificités obtenues par la comparaison avec le corpus Le Monde 2001

- La seconde méthode consistait à subdiviser le corpus d'informatique en six sous-corpus représentatifs (matériel, Internet, etc.; Voir le tableau 2) et à les comparer à l'ensemble des textes d'informatique. La liste des 25 premières spécificités positives obtenues par cette seconde méthode est reproduite dans le tableau 4.

Vingt-cinq premières SP+				
Unité	**Partie du discours**	**Fréquence brute**	**Valeur test**	**Sous-corpus**
internet	SBC	786	41.1712	SC2
mémoire	SBC	737	30.4842	SC1
option	SBC	1258	29.3401	SC6
site	SBC	306	28.9233	SC2
recherche	SBC	346	28.7659	SC2
imprimante	SBC	200	25.9019	SC4
papier	SBC	104	25.0858	SC4
fichier	SBC	2388	23.0704	SC6
implémentation	SBC	55	22.85	SC5
epson	SBP	44	22.101	SC4
recommander	ADJPAR	496	21.4903	SC6
algorithme	SBC	61	20.2627	SC5
synchronisation	SBC	63	20.0507	SC5
remplissage	SBC	38	19.994	SC5
noyau	SBC	438	19.9696	SC6
contrat	SBC	44	19.9632	SC5
laser	SBC	62	19.879	SC4
électronique	ADJ	202	19.8612	SC2
linux	SBP	422	19.6613	SC6
france	SBP	154	19.6566	SC2
étendue	SBC	75	19.1454	SC3
configuration	SBC	660	19.0063	SC6
imprimant	ADJ	52	18.9279	SC4
feuille	SBC	67	18.8884	SC4
réseau	SBC	292	18.7129	SC5

Tableau 4 : Listes des 25 premières spécificités obtenues par subdivision du corpus d'informatique en sous-corpus

Chaque méthode a produit des spécificités positives différentes, mais chacune contenait des unités potentiellement intéressantes par rapport aux objectifs que nous nous étions fixés. Une étude préalable (Lemay et al. en préparation) a comparé la liste de spécificités positives produites par les deux méthodes au contenu de deux dictionnaires commerciaux[2]. L'évaluation a porté sur les formes lexicales dont la première lettre était *A*, *C* et *P*[3].

Cette étude a révélé que la précision dans les deux méthodes se situe entre 45% et 55% (la variation est due au dictionnaire envisagé). Étonnamment, la différence entre la première méthode et la seconde n'est pas très élevée (entre 3% et 4%). Par ailleurs, l'étude a montré que la solution la plus intéressante consistait à combiner les listes de spécificités positives produites par les deux méthodes. Les taux de précision et de rappel de cette nouvelle liste sont reproduits dans le Tableau 5. Signalons que le rappel est calculé de deux manières. La première valeur (la moins élevée) présente le rappel en tenant compte de tous les termes répertoriés dans les dictionnaires[4]. La seconde valeur tient compte des termes qui apparaissent dans le corpus d'informatique. La combinaison des deux listes a comme conséquence d'améliorer considérablement le rappel.

Liste combinée de SP+				
	Nombre d'unités	**Nombre d'unités dans le dictionnaire**	**Précision**	**Rappel**
SP+	1021			
Ginguay	1,290 (854)	452	44,27 %	35,04 % (52,93 %)
Collin	1,311 (945)	512	50,20 %	39,05 % (54,78%)

Tableau 5 : Évaluation de la précision et du rappel de la liste combinée de spécificités lexicales

4. Application des critères lexico-sémantiques afin de valider les spécificités

La liste combinée contenant 1021 spécificités positives a été analysée afin de voir si les unités ramenées automatiquement correspondaient au type de termes à retenir dans le dictionnaire envisagé. Dans cette seconde évaluation, nous avons fait intervenir les critères lexico-sémantiques donnés à la section 2. Des exemples d'unités acceptées et rejetées sont donnés dans le tableau 6 (qui montre, de plus, le recoupement avec les dictionnaires utilisés dans la première évaluation et la méthode (1 ou 2) qui a généré la forme lexicale).

Les résultats de la sélection, quant à eux, ont été reproduits dans le tableau 7 et sont placés sous ceux obtenus à la suite de la comparaison avec les deux dictionnaires commerciaux. La précision ne peut être calculée en ce qui concerne l'application des critères lexico-sémantiques puisque le choix définitif des entrées n'est pas encore fixé.

Unité extraite	Partie du discours	Présent dans le Ginguay	Présent dans le Collin	Sélectionné	Méthode
abonné	SBC	oui	oui	oui	1 et 2
abonnement	SBC	non	non	oui	1 et 2
abonner	VB	non	non	oui	1
accès	SBC	oui	oui	oui	1 et 2
accessibilité	SBC	oui	non	oui	1
accessible	ADJ	oui	oui	oui	1
algorithme	SBC	oui	oui	oui	1 et 2
algorithmique	ADJ	non	non	oui	1 et 2
automatique	ADJ	oui	oui	oui	1
automatiquement	ADV	non	non	oui	1
automatisation	SBC	oui	oui	oui	1
caractère	SBC	oui	oui	oui	1 et 2
chemin	SBC	oui	oui	oui	2
clavier	SBC	oui	oui	oui	1 et 2
partage	SBC	oui	oui	oui	1 et 2
partageable	ADJ	oui	non	oui	1
partagé	ADJPAR	non	non	oui	1 et 2
aller	VB	oui	oui	non	2
absolument	ADV	non	non	non	2
accord	SBC	non	oui	non	1
affirmative	SBC	non	non	non	1
agir	VB	non	non	non	1
ainsi	ADV	non	non	non	1
alcool	SBC	non	oui	non	2
autre	ADJ	non	non	non	1
avantage	SBC	non	non	non	1
capable	ADJ	non	oui	non	1 et 2
centimètre	SBC	non	oui	non	2
chez	PREP	non	non	non	2
choix	SBC	non	oui	non	1
partir	VB.	non	non	non	1
permettre	VB	non	oui	non	1 et 2

Tableau 6 : Exemples d'unités retenues et rejetées par l'application de critères lexico-sémantiques

Liste combinée de SP+				
	Nombre d'unités	Nombre d'unités dans le dictionnaire	Précision	Rappel
SP+	1021			
Ginguay	1,290 (854)	452	44,27 %	35,04 % (52,93 %)
Collin	1,311 (945)	512	50,20 %	39,05 % (54,78%)
Critères lexico-sémantiques		553	54,16 %	

Tableau 7 : Spécificités retenues à la suite de l'application des critères lexico-sémantiques

Les résultats montrent que, sur le total des spécificités, 54,16 % ont été retenues en fonction des critères lexico-sémantiques énumérés à la section 2. Ces résultats sont à

peine plus élevés que ceux obtenus à la suite d'une comparaison de la liste de spécificités avec le contenu de deux dictionnaires d'informatique commerciaux. Ce résultat nous a quelque peu étonnée, puisque la seconde évaluation procédait à partir de la liste elle-même et non d'un travail lexicographique fait au préalable et totalement distinct du corpus utilisés pour générer les spécificités.

5. Conclusion

Ce travail montre que l'identification automatique des termes spécifiques par comparaison de corpus revêt un intérêt pour la sélection d'unités terminologiques susceptibles d'être retenues dans un dictionnaire spécialisé. Plus de la moitié des unités identifiées automatiquement ont été définies comme des candidats valables après l'application de critères lexico-sémantiques. En outre, ces premiers résultats sont comparables à ceux obtenus à la suite de la comparaison de la même liste de spécificités au contenu de deux dictionnaires commerciaux. La comparaison de corpus permet au terminographe d'ancrer ses choix préliminaires sur des données extraites de corpus, plutôt que sur des connaissances du domaine données a priori.

Toutefois, de nombreuses imperfections doivent être rectifiées. D'une manière générale, ces chiffres révèlent que même si les critères de sélection diffèrent d'un terminographe à l'autre, les taux de précision n'excéderont probablement pas 60 % compte tenu de la méthode et des corpus utilisés pour produire cette liste de spécificités.

Nous avons déjà évoqué le problème du bruit généré par les listes de spécificités. Une autre source de problèmes que nous n'avons pas abordé est celui du silence. Toutes les unités susceptibles d'être retenues comme entrée ne sont pas ramenées par la métrique appliquée et sont probablement distribuées dans les formes banales et spécificités négatives. Par exemple, les termes *automatique, automatiquement, automatisation* sont relevées comme étant des spécificités ; toutefois, *automatiser* n'est pas ordonné de cette manière. Ici encore, les problèmes devront être résolus par l'application d'une analyse terminographique.

Remerciements

Nous aimerions remercier les étudiants qui aidé dans le traitement des données : Sahara Iveth Carreño Cruz et Philippe Hanscom. Nous remercions également Anne-Laure Jousse et Chantal Lemay dont les travaux ont permis de parfaire les méthodes de sélection des termes d'informatique et Patrick Drouin qui a fait des adaptation à un programme qu'il a mis au point appelé TermoStat afin de le rendre apte à produire les spécificités recherchées ici.

Notes

[1] L'étiqueteur utilisé est celui de Brill et, plus spécificiquement, l'adaptation qui en a été faite dans Winbrill (ATILF 2003). Le lemmatiseur est FLEMM (Lecomte 1998; Namer 2000).

[2] Les dictionnaires en question sont le Collin (1996) et le Ginguay (1998).

[3] Ces lettres ont été choisies parce qu'elles contenaient le plus grand nombre de formes.

[4] Les dictionnaires contenaient des formes qui n'apparaissaient pas dans le corpus (par exemple, *abaque*).

Bibliographie

Ahmad, K. A. Davies, H. Fulford and M. Rogers. 1994. **"**What's in a Term ? The semi-automatic Extraction of Terms from Text", In Snell-Hornby, M. F. Pochhacker and K. Kaindl (eds.). Translation Studies. An Interdiscipline*,* Amsterdam/Philadephia: John Benjamins, pp. 267-278.

ATILF. 2003. Laboratoire ATILF (Analyse et traitement informatique de la langue française (page consultée en mai 2002 : http://www.atilf.fr/)

Binon, J., S. Verlinde, J. Van Dyck et A. Bertels. 2000. Dictionnaire d'apprentissage du français des affaires. Dictionnaire de compréhension et de production de la langue des affaires, Paris : Didier.

Chung, T. M. 2003. "A Corpus Comparison Approach for Terminology Extraction", Terminology 9(2), pp. 221-246.

Collin, S.M.H., F. Laurendeau and B. Mouget. 1996. Le Bilingue de l'informatique : dictionnaire français-anglais, anglais-français*.* Coll. « Peter Collins », Middlesex: Peter Collin.

Dancette, J. et C. Réthoré. 2000. Dictionnaire analytique de la distribution. Analytical Dictionary of Retailing, Montréal : Les Presses de l'Université de Montréal.

Descamps, J.L. 1976. Dictionnaire contextuel de français pour la géologie : essai de classement d'un concordance de français scientifique et étude critique, Paris : Didier.

Drouin, P. 2003. "Term Extraction Using Non-technical Corpora as a Point of Leverage", Terminology 9(1), pp. 99-115.

Frawley, W. 1988. "New forms of Specialized Dictionaries", International Journal of Lexicography 1(3), pp. 189-213.

Ginguay, M. 1998. Dictionnaire français-anglais d'informatique : bureautique, télématique, micro-informatique, 6^{e} éd., 2^{e} tirage avec mise à jour, Paris : InterEditions.

Jousse, A.L. et M. Bouveret. 2003. "Lexical functions to represent derivational relations in specialized dictionaries", Terminology 9(1), pp. 71-98.

Kilgariff, A. 2001. "Comparing Corpora". *International Journal of Corpus Linguistics* 6(1), pp. 1-37.

L'Homme, M.C. 1998. « Définition du statut du verbe en langue de spécialité et sa description lexicographique », Cahiers de lexicologie 73(2), pp. 61-84.

L'Homme, M.C. 2003. "Capturing the lexical structure in special subject fields with verbs and verbal derivatives: A model for specialized lexicography", International Journal of Lexicography 16(4), pp. 403-422.

Lafon, P. 1980. "Sur la variabilité de la fréquence des formes dans un corpus", Mots 1, pp. 128-165.

Lebart, L. and A. Salem. 1994. Statistique textuelle*,* Paris : Dunod.

Lecomte, J. 1998. Le catégoriseur Brill14-JL5 / WinBrill-0.3, INaLF/CNRS (page consultée le 13 février 2003 http://www.atilf.fr/winbrill/BRILL14-JL5_WinBrill.doc

Lemay, C., M.C. L'Homme and P. Drouin. en préparation. "Two Methods for Extracting "Specific" Single Word Terms (SWTs) from Specialized Corpora : Experimentation and Evaluation".

Mel'čuk, I. et al. 1984-1999. *D*ictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques *1*, Montréal : Les Presses de l'Université de Montréal.

Mel'čuk, I., A. Clas, A. et A. Polguère. 1995. Introduction à la lexicologie explicative et combinatoire, Louvain-la-Neuve (Belgique) : Duculot / Aupelf - UREF.

Namer, F. 2000. "FLEMM: A Rule-based Morphological Analyzer of Inflected French Words". *T.A.L.* 41(2), pp. 523-548.

Polguère, A. (2003). « Collocations et functions lexicales : pour un modèle d'apprentissage », In Grossmann, F. et A. Tutin (éd.). Les collocations. Analyse et traitement, Coll. Travaux et recherches en linguistiques appliquée, Paris : Éditions de Werelt.

Rayson, P. and R. Garside. 2000. "Comparing Copora Using Frequency Profiling", In *Proceedings of the Workshop on Comparing Corpora, 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, pp. 1-6.