Extraction semi-automatique des néologismes dans la terminologie du terrorisme

Patrick Drouin[1], Annie Paquin[1], Nathan Ménard[2]

[1] OLST / ÉCLECTIK, Département de linguistique et de traduction, Université de Montréal
C.P. 6128, Succursale Centre-Ville, Montréal (Québec) H3C 3J7, Canada
patrick.drouin@umontreal.ca, annie.paquin@umontreal.ca

[2] GRELT, Département de linguistique et de traduction, Université de Montréal
C.P. 6128, Succursale Centre-Ville, Montréal (Québec) H3C 3J7, Canada
nathan.menard@umontreal.ca

Abstract

In this paper, we present a semi-automatic method to extract neologisms. Our technique uses a specialized corpus relating to the field of terrorism subdivided in two subcorpora spread over two distinct periods as well as a non-specialized corpus. We will compare these corpora using the term extractor *TermoStat* to find the lexical specificities of each corpus and study them under a diachronic angle to identify the neologisms.

Résumé

Dans cet article, nous explorerons une méthode d'extraction semi-automatique permettant de repérer les néologismes terminologiques. Notre technique de recherche consiste en la constitution d'un corpus spécialisé dans le domaine du terrorisme subdivisé en deux sous-corpus s'échelonnant sur deux périodes distinctes ; ainsi que d'un corpus de référence relevant de la langue générale. Nous effectuerons une série de comparaisons entre ces différents corpus et sous-corpus à l'aide de l'extracteur terminologique *TermoStat* afin de relever les spécificités lexicales de chacun des sous-corpus pour ensuite les évaluer sous un angle diachronique et ainsi en repérer les néologismes.

Mots-clés : terminologie, néologie, veille terminologique, acquisition automatique des termes, lexicométrie, terrorisme.

1. Introduction

Les développements des dernières années en acquisition automatique des termes ont permis de profondes mutations et une accélération considérable du dépouillement des corpus spécialisés afin d'en extraire la terminologie. Nous croyons qu'il est possible avec une méthodologie semblable, de rechercher plus spécifiquement les néologismes dans un domaine donné à l'aide de vastes corpus. Les néologismes mettent généralement un certain temps avant d'être décelés et inclus dans les dictionnaires spécialisés. Cette lacune n'est pas uniquement due aux processus de normalisation et de sélection des nouveaux termes, mais également à l'ampleur du traitement des documents. Pourtant, l'étude des néologismes comporte de nombreuses applications (Lerat, 1990). Outre l'enrichissement des dictionnaires spécialisés, qui servent de transition à leur intégration dans les dictionnaires généraux, ainsi que les thésaurus et les glossaires ; une meilleure connaissance des néologismes permet de mieux observer les mécanismes de création lexicale, de connaître les variantes d'un terme en processus d'intégration et de faciliter la normalisation (Humbley, 2003).

Nous estimons que le logiciel *TermoStat* (Drouin 2003), qui facilite le recensement des éléments terminologiques dans les textes spécialisés en procédant à une mise en opposition de corpus, peut tout aussi bien être utilisé afin de repérer les néologismes. Afin d'y parvenir, les corpus exploités divergent puisqu'ils sont issus de tranches synchroniques différentes et qu'ils correspondent à des niveaux de spécialisation différents. Pour la présente étude, nous avons constitué un corpus de textes spécialisés reliés au terrorisme. Puisque ce domaine a été fortement couvert au cours des dernières années par les médias, surtout depuis les attentats du 11 septembre 2001, que la lutte contre le terrorisme est devenue une priorité mondiale pour la plupart des pays occidentaux, et que la perception envers celui-ci semble avoir subi d'importantes mutations, nous présupposons certains changements observables dans les textes portant sur ce domaine qui pourraient se traduire par un nombre important d'innovations lexicales.

La section qui suit décrira brièvement quelques projets qui ont été réalisés en extraction assistée par ordinateur des néologismes. Nous exposerons par la suite la méthodologie que nous avons utilisée pour la constitution du corpus, les paramètres d'acquisition des données par *TermoStat* et l'analyse des résultats. Nous observerons finalement les résultats à partir desquels nous pourrons envisager des applications de notre méthode à l'étude des néologismes.

2. Travaux antérieurs

La recherche de néologismes a longtemps été une discipline essentiellement manuelle. Avec l'essor du traitement automatique de la langue (TAL) et en raison du nombre important de documents spécialisés à étudier afin de garantir un résultat satisfaisant, plusieurs études ont été entreprises dans les années 90 afin d'évaluer les perspectives d'extraction assistée par ordinateur des néologismes et de permettre une veille terminologique.

Ces études se basent pour la plupart sur une comparaison des unités recensées avec un corpus d'exclusion, qui est généralement un ouvrage de référence. C'est le cas, de Mathieu, Gross et Fouquéré (1998) qui, pour repérer les néologismes dans la langue générale, ont créé un corpus composé d'articles du quotidien *Le Monde*, qu'ils ont comparé au *Trésor de la Langue Française* afin de repérer les formes qui n'étaient pas répertoriées dans ce dernier. C'est aussi la méthode qu'ont adopté Cabré *et al.* (2003) pour l'espagnol et le catalan grâce au logiciel *Sextan* qui relève toutes les unités lexicales dans un corpus de journaux qui ne sont pas reconnues par le dictionnaire électronique de l'outil. Des expériences semblables ont aussi été menées en terminologie par Jaccarini (1999) ainsi que L'Homme *et al.* (1999) qui ont pour leur part utilisé la *Banque de terminologie du Québec* (BTQ), aujourd'hui le *Grand dictionnaire terminologique* de l'Office québécois de la langue française, comme point de référence.

Certaines méthodes sont secondées par un système de filtres visant à retenir les néologismes potentiels et à éliminer les unités lexicales non pertinentes. Ainsi, le système de détection semi-automatique des néologismes dans les domaines spécialisé, *Cenit* (*Corpus-based English Neologism Identifer Tool*), de Roche et Bowker (1999) traite les candidats en les comparant d'abord à un dictionnaire général puis à des dictionnaires spécialisés, pour ensuite les soumettre à une série de filtres. Le premier filtre élimine les dérivés morphologiques des unités lexicales présentes dans les dictionnaires. Cependant, puisque la dérivation est un processus très répandu en création lexicale et qu'il pourrait par conséquent éliminer un nombre considérable de néologismes, ce filtre est facultatif. Les filtres suivants éliminent les

noms propres ainsi que les sigles décrivant des produits ou des sociétés. Le système dresse finalement une liste des termes retenus qui sera par la suite triée par un terminologue.

L'emprunt est lui aussi une forme de néologisme très répandu, Sablayrolles (2000) et Jacquet-Pfau (2003) souligne l'importance de le repérer adéquatement en TAL. À cet effet, elle suggère une reconnaissance des emprunts par des critères graphiques, d'organisation et de combinaison des graphèmes ainsi que par leur structure morpho-syllabique.

Une méthode d'extraction semi-automatique des néologismes se basant sur une approche statistique a été réalisée par Janicijevic et Walker (1997). Leur outil, appelé *NeoloSearch*, traite des corpus de textes recueillis sur Internet et identifie les néologismes à l'aide d'un test statistique (*z-score*) reposant sur des observations de fréquence. Les noms propres, les mots rares et ceux contenant des erreurs typographiques sont ensuite éliminés pour créer une liste de néologismes potentiels. Les contextes et les modes d'affixation les plus courants de ceux-ci sont répertoriés afin d'en étudier les modes de création.

Dans cet article, nous proposons une démarche qui se distingue aussi de celles précédemment mentionnées puisque nous ne faisons appel à aucune ressource linguistique extérieure (dictionnaire, thésaurus, banque de terminologie). Nous nous rapprochons sensiblement de la démarche proposée par Janicijevic et Walker (1997), puisqu'elle repose elle aussi sur une observation des variations de fréquence en corpus. Cependant, nous adoptons une mesure statistique issue des travaux de Lafon (1980) et de Lebart et Salem (1994), le calcul des spécificités.

3. Méthodologie

3.1. Description des corpus

La méthode que nous proposons repose sur une analyse statistique, qui s'appuie sur le calcul des spécificités (Lafon 1980) tel qu'implanté au sein du logiciel *TermoStat* (Drouin 2004). Ce dernier met en opposition deux corpus de nature différente de façon à faire ressortir leurs particularités lexicales.

Pour la présente recherche, nous utilisons un corpus de langue journalistique construit à partir de l'ensemble des articles publiés au cours de l'année 2002 dans le journal *Le Monde* qui totalise environ 30 millions de mots. Ce corpus sera noté *MONDE*. Ces articles sont rédigés dans une langue non spécialisée et portent sur des sujets très diversifiés. Cependant, puisque le terrorisme était un sujet fortement exploité par les media à cette période, nous avons cru nécessaire de vérifier quelle proportion des articles traitait de cette thématique. Nous en avons relevé 565 sur un total de 25 280, ce qui représente 2,24% des articles. Nous croyons que cette proportion n'est pas excessive et qu'elle ne devrait pas influencer les résultats, d'autant plus que ces articles ne correspondent pas du même niveau de spécialisation que le corpus d'analyse.

Le second corpus, que nous appellerons *TERROR*, est constitué de textes spécialisés portant sur le terrorisme dont la majorité provient de sites Internet gouvernementaux ; de revues de politique internationale et militaires ; de sites Internet spécialisés sur le terrorisme, et d'instituts universitaires de criminologie ou de sciences politiques. Une plus petite partie du corpus est issue de documents numérisés à partir d'encyclopédies du terrorisme et d'articles de revues spécialisées. La date de publication de ces documents s'échelonne sur une période de dix ans, soit de 1995 à 2005.

Pour les fins de la présente recherche, le corpus *TERROR* sera parfois subdivisé en deux sous-corpus : le sous-corpus *TERROR-2001* comporte les documents publiés avant le 11 septembre 2001 alors que le sous-corpus *TERROR+2001* regroupe ceux ayant été publiés subséquemment. La taille du *TERROR-2001* est de 502 659 occurrences et celle du *TERROR+2001* est de 510 022 occurrences pour un total de 1 012 681 occurrences pour le corpus *TERROR*. Les documents composant le corpus d'analyse sont en français[1].

Le domaine du terrorisme est relié à de nombreuses disciplines telles la criminologie, les sciences politiques, la sociologie, la stratégie militaire, voire même les sciences pures dans le cas du bioterrorisme, du terrorisme chimique et du terrorisme nucléaire. C'est la raison pour laquelle nous avons cherché à équilibrer le nombre d'occurrences issues de ces domaines dans chaque sous-corpus tel que l'illustre le **Tableau 1.**

Sous-domaine	Avant 11 septembre	Après 11 septembre
Bioterrorisme / terrorisme chimique / nucléaire	40 824	56 501
Causes du terrorisme	7 920	11 264
Cyberterrorisme	6 919	6 062
Définition du terrorisme	44 251	45 155
Droit du terrorisme	3 881	12 529
Histoire et évolution du terrorisme	115 613	92 962
Lutte au terrorisme	86 302	117 482
Organisations terroristes	62 439	65 092
Techniques du terrorisme	31 062	31 155
Terrorisme aérien	29 801	17 872
Terrorisme islamique	40 952	27 756
Terrorisme national	4 244	10 853
Terrorisme religieux	28 451	15 339
TOTAL DES OCCURENCES	502 659	510 022

Tableau 1 : Répartition des domaines du corpus d'analyse (nombre d'occurrences)

3.2. Description des extractions

Nous avons soumis les sous-corpus au logiciel *TermoStat* (Drouin, 2003) qui, après avoir étiqueté et lemmatisé toutes les unités lexicales des corpus, procède à l'identification des spécificités (Drouin, 2004) du corpus d'analyse. Afin de tester les différentes possibilités offertes par cette méthode en analysant les résultats obtenus par les diverses combinaisons possibles entre les sous-corpus et le corpus de référence ; cinq extractions différentes ont été réalisées (**Tableau 2**). Nous avons d'abord opposé le sous-corpus *TERROR-2001*, puis en second lieu le sous-corpus *TERROR+2001* au corpus *MONDE* afin de relever les termes de la langue de spécialité les plus représentatifs par rapport à la langue non spécialisée pour les deux périodes étudiées. Nous avons ensuite opposé chacun des sous-corpus l'un à l'autre pour extraire les termes dont la fréquence varie de manière significative dans une perspective exclusivement diachronique. Finalement, nous avons comparé le corpus *MONDE* au corpus *TERROR* dans son intégralité.

[1] Une partie des documents du corpus *TERROR* est issue d'un processus de traduction depuis l'anglais. En effet, c'est dans cette langue que sont rédigées la majorité des études de ce domaine. Il pourrait être intéressant d'étudier les néologismes issus des textes traduits afin de voir s'ils constituent des variantes des formes repérés dans les textes rédigés initialement en français tel que l'avancent Hermans et Vansteelandt (1999).

3.3. Description de l'analyse des résultats

Le nombre de candidat termes (CT) retenus par chacune de ces extractions varie selon le corpus de référence utilisé tel que le démontre le **Tableau 2**. Ainsi, les extractions pour lesquelles le corpus de référence est le corpus *Monde* comptent chacune plusieurs dizaines de milliers de candidats termes. Nous avons décidé de ne traiter que les 5000 candidats termes dont les poids étaient les plus élevés.

Extraction	Point de comparaison	Corpus comparé	Nb de CT extraits
1	*MONDE*	*TERROR-2001*	38 447
2	*MONDE*	*TERROR+2001*	37 087
3	*TERROR+2001*	*TERROR-2001*	186
4	*TERROR-2001*	*TERROR+2001*	448
5	*MONDE*	*TERROR*	16 675

Tableau 2 : Nombre de candidats termes extraits par TermoStat

Dans le but de comparer les résultats des candidats termes entre chacune des extractions, nous leur avons attribué un rang. Ce rang est déterminé selon l'ordre décroissant du poids que le candidat terme a obtenu lors de l'extraction. En effet, suite au calcul des spécificités, le logiciel attribue la valeur issue de ce calcul aux CT. De cette manière, le candidat dont la spécificité est la plus élevée occupe le rang 1, le deuxième plus élevé, le rang 2, etc. C'est en étudiant les variations des rangs des termes de cette liste lors des divers types d'extractions que seront identifiés les néologismes.

Plusieurs candidats termes s'étant retrouvés dans plus d'une extraction, une comparaison a été faite entre les listes issues des cinq extractions afin d'éviter les répétitions. Cette phase d'épuration manuelle de la liste a permis de réduire le total à 9631 CT. Nous avons par la suite supprimé tous les noms propres de la liste ainsi que les CT faisant explicitement référence à une zone géographique ou un groupe particulier, tels que *bloc soviétique, attentat Sikh, brigade révolutionnaire corse,* ou *bombe de Oklahoma*. Ces unités lexicales renvoient à des entités ou des événements trop précis pour être analysé comme des termes du domaine. Il serait intéressant, afin de réduire le temps d'analyse lors d'expérimentations futures, de procéder à une élimination automatique de ces entrées de la liste grâce à un outil d'identification automatique des entités nommées (Poibeau, 2005).

Nous avons ensuite procédé à la sélection des termes du terrorisme. Cette sélection s'est faite, dans un premier temps, selon les critères de sélection de termes de L'Homme (2004). Nous avons donc considéré les CT dont le sens était relié au terrorisme ; nous avons également étudié les CT qui entretiennent des relations actancielles avec d'autres termes du terrorisme afin de vérifier s'ils constituent eux-mêmes des termes ; nous avons finalement observé les CT présentent des relations morphologiques ou paradigmatiques avec d'autres termes du terrorisme.

Malgré l'application de tels critères de sélection, le statut de certaines unités nous semblait problématique. Comme nous l'avons dit précédemment, le terrorisme touche à de nombreux domaines et nous avons retrouvé dans la liste un grand nombre de termes de domaines connexes sans que ceux-ci ne soient liés très concrètement au terrorisme. Nous avons donc, dans un deuxième temps, effectué une seconde sélection à caractère exclusivement sémantique qui a permis de retrancher les termes qui sont déjà bien intégrés dans d'autres domaines. Pour ce faire, nous avons dressé une liste des sous-domaines que nous désirons

analyser avec pour objectif de restreindre l'étendue disciplinaire des termes étudiés : *techniques du terrorisme* (armes, logistique, types de terrorisme, actions, etc.), *acteurs* (type et structure des groupes terroristes, cibles humaines, etc.), *lutte au terrorisme* (équipement, stratégies, législation, etc.) et *idéologie* (motivations des terroristes, doctrines, etc.). Une dernière étape d'épuration des résultats avait pour but d'éliminer les termes déjà bien intégrés dans les domaines connexes, plus particulièrement les termes reliés aux armes chimiques, tel que *gaz neurotoxique, charbon bactérien* ou *organisme pathogène*. Suite à l'imposition de ces dernières contraintes, nous avons considérablement réduit la liste en l'abaissant à moins mille termes.

Nous avons ensuite étudié les contextes des termes exclusifs aux 2^{e} et 4^{e} extractions à l'aide du logiciel SATO (Duchastel *et al*., 2004). En annotant adéquatement les fichiers du corpus, cet outil nous a permis d'observer l'évolution détaillée des termes dans le corpus en fonction des années de publication des documents et des auteurs. Nous avons ainsi pu déterminer si un néologisme potentiel compte des attestations antérieures au 11 septembre 2001 et s'il est utilisé par plus d'un auteur. Le logiciel SATO permet donc de compléter les travaux initiaux effectués à l'aide de *TermoStat*.

Puisque l'étendue chronologique du corpus est tout de même limitée, nous avons tenu à vérifier que les unités lexicales identifiées comme des néologismes potentiels n'apparaissaient pas dans un corpus de presse plus important. Pour ce faire, nous avons eu recours au service *Biblio Branchée* de la société Cedrom-SNI qui permet d'interroger un corpus de presse couvrant environ une vingtaine d'années. Afin de compléter ce processus, les banques de terminologie du gouvernement du Canada (*Termium Plus*) et du gouvernement du Québec (*Le grand dictionnaire terminologique*) ont été exploitées afin de valider la terminologie plus technique recensée dans notre corpus.

4. Résultats

C'est au sein de la cette liste de termes que nous avons cherché les néologismes potentiels en comparant les termes issus des différentes extractions. L'analyse des résultats nous permet de constater un nombre considérable de néologismes de formes différentes issus de processus de création variés tel que l'illustrent les quelques exemples du **Tableau 3**. Dans la présente section, nous observerons d'abord les différentes formes de néologismes repérés à l'aide de la méthodologie décrite précédemment. Nous présentons ensuite d'autres observations faites sur les résultats obtenus et sur leur utilité potentielle en lexicologie et en terminologie.

4.1. Néologie morphologique

Sous l'appellation néologie morphologique, ou néologie de forme, nous regroupons à la fois les phénomènes reliés aux processus de dérivation et de formation syntagmatique. On note qu'à partir de préfixes productifs dans la langue générale sont créés de nombreuses formes dont *agro-* (*agroterrorisme*) et *hyper* (*hyperterrorisme*). On observe également une quantité importante de dérivations par ajout d'un suffixe comme *–i*sme et –iste ayant comme radical un emprunt (*djihadisme, djihadiste, jihadisme, jihadiste, karyanistes, takfiriste)*. Le processus de dérivation touche aussi les noms propres qui exploitent à leur tour le suffixe *–isme* (*benladenisme*) et le suffixe *–logue* (*qaïdologues*).

Les résultats des différentes extractions contiennent un nombre important d'unités syntagmatiques nouvelles. Il peut être surprenant de constater un nombre si élevé de ce type de néologismes, mais nous avons cru pertinent de recenser toutes les nouvelles combinaisons de termes présentes dans le corpus, tels que *arme asymétrique, argent terroriste, arme de*

terreur de masse afin de rendre compte de tous les nouveaux cooccurrents possibles. Nous considérons donc que l'unité nominale *arme asymétrique* et l'adjectif qui lui confère son statut de néologisme *asymétrique,* doivent être recensées séparément. Ce dernier entre d'ailleurs dans la création de plusieurs unités intéressantes dont *conflit asymétrique*, *guerre asymétrique, menace asymétrique,* etc.

Le dernier phénomène observé est la création d'un nouveau sigle *ADM* qui est utilisé comme substitut au néologisme syntagmatique *armes de destruction massive*. Cette présence plus que timide d'acronymes ou de sigles dans les résultats obtenus est surprenante puisque ces unités sont souvent très présentes dans les analyses sur corpus.

Néologismes	**Extraction**				
	1	**2**	**3**	**4**	**5**
ADM		59		113	186
argent terroriste		2553			
arme asymétrique		2572			
arme de terreur de masse		2572			
asymétrie		168		97	527
agroterrorisme		1010			2390
benladenisme		599			1463
conflit asymétrique		2162			
djihadisme		3354			
djihadistes		873			2168
guerre asymétrique		465			1024
hyperterrorisme		684			1724
jihadis		1691			3959
jihadisme		306			821
jihadiste		86		135	265
karyanistes		3912			
menace asymétrique		335			788
nébuleuse		1346			3052
qaïdologues		4589			
réseautique		590			1447
takfiriste		272			737

Tableau 3. Présentation de néologismes

4.2. Néologie sémantique

Ce phénomène, aussi nommé néologie de sens, consiste à réutiliser du matériel lexical existant en lui assignant une nouvelle signification. On assiste bien souvent à la récupération, en vue de désigner un nouveau concept, d'une forme utilisée d'un domaine bien établi ou encore de la langue générale. Au sein de notre corpus, nous avons repéré quelques néologismes sémantiques : *asymétrie, nébuleuse* et *réseautique*. Les néologismes sémantiques sont généralement peu fréquents dans les domaines techniques, nous ne sommes donc pas étonnés de leur importance limitée dans les résultats obtenus.

4.3. Néologie d'emprunt

Plusieurs emprunts ont été extraits par le logiciel *TermoStat*, mais aucun de ces CT ne constituait à la fois un terme du terrorisme et un néologisme. Nous avons relevé un nombre important d'emprunts déjà largement utilisés en 2001 ainsi que quelques emprunts récents tels que *karyan*, *mufti* et *takfir*. Ces derniers sont toutefois rattachés à des domaines qui s'éloignent trop de celui du terrorisme pour que l'on puisse les traiter comme des termes. Il est cependant important de souligner que ceci ne les empêche pas d'être dérivés pour créer des néologismes morphologiques qui sont pour leur part directement liés au terrorisme. Nous avons d'ailleurs signalé *takfiriste* et *karyaniste* dans une section précédente.

4.4. Autres observations

L'utilisation d'une technique comme celle proposée dans cet article offre des pistes intéressantes pour le travail terminologique et lexicographique. En effet, elle permet de dépister et d'observer des variantes terminologiques comme *djihadisme / jihadisme* et *djihadistes / jihadiste / jihadis* qui illustrent une certaine incertitude du point de vue de la graphie à adopter. Les organismes officiels désireux de documenter ou d'influencer l'usage pourront alors y trouver des informations précieuses.

Dans une perspective semblable, il est possible d'étudier les termes qui sont intégrés dans la langue depuis un certain temps, mais dont l'intégration orthographique n'est pas tout à fait complétée. Comme l'illustre le **Tableau 4**, les emprunts *jihad* et *moudjahidine*, même s'ils étaient fréquemment utilisés dans le sous-corpus *TERROR-2001*, présentent de nombreuses variantes orthographiques. On peut cependant constater que l'orthographe de *jihad* semble se stabiliser dans le sous-corpus *TERROR+2001*. Ainsi, en plus de participer à l'étape de dépouillement des néologismes (Célestin, 2003), cette méthode peut également contribuer à faire un suivi sur leur implantation dans la langue de spécialité.

Variantes terminologiques	**Extraction**				
	1	**2**	**3**	**4**	**5**
djihad	111	257			102
djihâd	1141				2538
jihad	189	7		7	19
jihâd	87		98		207
moudjahiddines		1741			4072
moudjahidin	119	1297			220
moudjahidine	636				1423
mudjahidin	4159				
mujahideen	1226	4198			1542
mujahidin		968			1816

Tableau 4. Présentation de variantes terminologiques

L'observation de ces variations de fréquence permet d'inclure dans les ouvrages de références des néologismes qui existaient déjà mais dont la fréquence d'utilisation n'était pas suffisamment élevée pour être pris en compte. En effet, pour des raisons économiques, tous les néologismes ne peuvent être décrits dans les dictionnaires (Humbley, 1993), il est donc nécessaire d'étudier l'évolution de leur fréquence sur une base diachronique. La

méthodologie décrite ici permet donc de constater et documenter l'émergence de certaines formes.

5. Conclusion

L'intégration de l'approche documentée ici dans un atelier plus large incluant divers outils d'analyse de texte (SATO, Duchastel *et al*., 2004 par exemple) et de constitution continue de corpus à partir du Web ou de fils de presse (BootCaT Tools, Baroni et Bernardini, 2004 ; TerminoWeb, Agbago et Barrière, 2005) pourrait donner naissance à un environnement de travail du *néologue*. L'intérêt d'une alimentation continue du corpus de référence est sans équivoque puisqu'elle permet de dépister sur une base quotidienne les nouvelles graphies ou, sur une période de temps légèrement plus longue, l'apparition potentielle de nouveaux concepts. Un tel environnement de travail saurait facilement trouver sa place auprès d'un public intéressé à l'évolution des thématiques et des concepts dans divers domaines sociopolitiques. Dans le domaine plus strict du terrorisme, les analystes militaires pourraient ainsi obtenir des clés d'accès rapide à la masse documentaire qu'ils ont à analyser et gérer sur une base quotidienne.

Notre méthode, qui se base sur la comparaison de corpus appartenant à des périodes et des niveaux de spécialisation distincts, permet donc de repérer des néologismes en comparant différentes extractions réalisées grâce au calcul des spécificités. Nous avons constaté que cette méthode permet de répertorier plusieurs types de néologismes tout en permettant d'étudier l'évolution de leur fréquence et de leurs variantes orthographiques. Il en demeure toutefois que, pour en arriver à ces résultats, le terminologue doit étudier plusieurs milliers de candidats termes issus des multiples extractions. Nous prévoyons donc poursuivre l'exploration de cette méthode et la rendre encore plus efficace en tentant de réduire ce dépouillement manuel.

Références

Agbago A. et C. Barrière (à paraître). Corpus construction for terminology. *Proceedings from the Corpus Linguistics Conference Series*, 1(1).

Cabré M.T. *et al.* (2003). L'observatoire de néologie : conception, méthodologie, résultats et nouveaux travaux. In Sablayrolles J.F. (dir.) *L'innovation lexicale*, Champion, Paris : 125-147.

Celestin T. *et al.* (2003). Le phénomène de la néologie technique et scientifique au Québec - Bilan et perspectives. Relaiter. Colloque international Rome <http://www.realiter.net/roma/celestin.htm> (visité le 1er novembre 2005).

Drouin P. (2004). Spécificité lexicales et acquisition de la terminologie. In *Le poids des mots : Actes des JADT 2004* : 355-352.

Drouin P. (2003). Term Extraction Using non-Technical Corpora as Point of Leverage. *Terminology*, vol (9/1) : 99-115.

Duchastel J., Daoust F. et della Faille D. (2004). SATO-XML : une plateforme Internet ouverte pour l'analyse de texte assistée par ordinateur In *Le poids des mots : Actes des JADT 2004* : 353-363.

Hermans H et Vansteelandt A (1999). Néologie traductive. *Terminologies nouvelles*, n° 20 : 37-43.

Humbley J. (2003). La néologie en terminologie. In Sablayrolles J.F. (dir.) *L'innovation lexicale*. Champion. Paris : 260-278.

Humbley J. (1993). L'observation de la néologie terminologique : l'expérience du CTN. *La Banque des mots* : 65-73.

Jaccarini A. (1999). Utilisation d'une banque de textes en terminographie. *Terminologies nouvelles*, n°20 : 17-24.

Jacquet-Pfau C. (2003). Du statut de l'emprunt en traitement automatique des langues. In Sablayrolles J.F. (dir.) *L'innovation lexicale*. Champion, Paris : 79-97.

Janicijevic T. and Walker D (1997). NeoloSearch : Automatic detection of neologisms in French Internet documents. ACHALLC' 97, Kingston.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, vol. (1) : 128-165.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Denud.

Lerat P. (1990). Sélection et analyse de termes nouveaux dans une base de données prédictionnairiques. *Cahiers de lexicologie* : 256-260.

L'Homme M-C. (2004) *La terminologie : principes et techniques.* Les Presses de l'Université de Montréal : 64-66.

L'Homme M-C., Bodson C. et Valente R. (1999). Recherche terminographique semi-automatisée en veille terminologique : expérimentation dans le domaine médical. *Terminologies nouvelles*, n°20 : 25-35.

Mathieu Y-Y., Gross G., Fouqueré C., (1998). Vers une extraction automatique des néologismes. *Cahier de lexicologie*, n°72 : 199-208.

Poibeau T. (2005). Sur le statut référentiel des entités nommées. *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2005)*.

Roche S. Bowker, L. (1999). *Cenit :* Système de détection semi-automatique des néologismes. *Terminologies nouvelles*, n°20 : 12-15.

Sablayrolles J.F. (2000). *La néologie en français contemporain : examen du concept et analyse de productions néologique récentes*. Champion, Paris.