Cenit: Système de détection semi-automatique des néologismes

Les langues évoluent à un rythme accéléré, donc la détection manuelle des néologismes devient encore moins pratique. Cet article décrit *Cenit*, un système de veille néologique qui vise à identifier des néologismes dans les domaines spécialisés. Le système utilise une stratégie fondée sur l'exploitation de corpus. Nous exposons les composants et les divers stades de traitement.

Termes clés: néologismes; système de détection des néologismes; corpus; filtres.

1 Introduction

D'après Le Meur et Depecker (1995: 48): «Les langues évoluent à la fin du XXe siècle de façon prodigieuse, sous la poussée notamment des nouvelles techniques et de la révolution informationnelle en cours». Pour le terminologue au travail, il n'est plus raisonnable de travailler sans outils, donc l'application des ordinateurs aux divers aspects du travail terminologique continue à gagner en popularité. Depuis longtemps, les ordinateurs sont employés pour le stockage et l'organisation des données. Plus récemment, les techniques d'intelligence artificielle ont été adoptées en terminologie pour créer une nouvelle génération de ressources terminologiques qu'on appelle base de connaissances terminologiques (BCT) (Miller *et al.* 1991; Condamines 1995). Les BCT permettent non seulement le stockage des données linguistiques, mais de plus ils fournissent au terminologue la possibilité de saisir et de modéliser les données notionnelles et les arborescences de domaine.

En ce qui concerne les applications plus automatisées, des chercheurs ont examiné la possibilité de l'extraction automatique de termes, et des approches et des projets différents sont récapitulés dans Kageura et Umino (1996) ainsi que dans Bourigault *et al.* (1998). De même, plusieurs chercheurs commencent à étudier la possibilité d'informatiser la détection des néologismes, travail qui prend du temps et qui nécessite une main-d'œuvre importante si on le fait manuellement. Il existe déjà quelques systèmes créés dans le but de détecter des néologismes. Le système *Aviator* (Blackwell 1993; Collier 1993; Renouf 1993) identifie les néologismes dans la presse anglaise pour fournir aux enseignants des ressources lexicales. De même, le projet *Obneb* (Cabré et de Yzaguirre 1995) a pour but la détection des néologismes dans la presse, mais cette fois en espagnol et en catalan. L'outil *NeoloSearch* (Janicijevic et Walker 1997) identifie les néologismes français qui se trouvent sur l'Internet, tandis que le système *Cordon* (W3 *Cordon*) utilise des techniques statistiques et donc n'est pas lié à une seule langue. Dans cet article, nous décrivons *Cenit* (*Corpus-based English Neologism Identifier Tool*), un prototype mis au point à Dublin City University. *Cenit* diffère des systèmes susmentionnés car il s'adresse spécifiquement aux terminologues et donc, contrairement à l'*Obneb* par exemple, *Cenit* détecte des néologismes dans les domaines spécialisés. L'*Obneb*, par contre, ne sert pas les besoins des terminologues car, selon Cabré et de Yzaguirre (1995: 90), l'*Obneb* «exclut préalablement toute nouveauté terminologique qui appartient à un domaine spécifique de la technologie, de la science ou de l'industrie».

2 Néologismes

Avant de décrire en détail le prototype, il faut s'arrêter un bref instant sur la question: «Qu'est-ce qu'un néologisme?» Nous sommes d'accord avec Rey (1995: 63) qui note que l'établissement d'une définition précise de «néologisme» est une tâche assez difficile. Néanmoins, la plupart des terminologues s'accordent sur les points suivants: un néologisme est une *unité lexicale* qui a l'air *innovateur* dans la langue à l'étude (Dubuc 1985: 111; Rey 1995: 64; Rondeau 1984: 122; Sager 1990: 79). Mais ils ne s'accordent pas toujours sur la nature spécifique de l'unité lexicale ou le genre particulier de la nouveauté. Par exemple, Rondeau (1984: 127) identifie trois grandes catégories de néologismes (néologisme de forme, néologisme de sens, et emprunt), tandis que Rey (1995: 68-70) précise trois catégories différentes (néologisme formel (y compris les emprunts), néologisme sémantique et néologisme pragmatique). Dubuc (1985: 113-4), pour sa part, précise deux catégories principales: les néologismes de formation dite indirecte (où l'on donne à un mot existant un sens nouveau) et ceux de formation directe (où l'on crée une nouvelle entité lexicale, soit de toutes pièces, soit par regroupement d'éléments existants). Finalement, Sager (1990: 71) n'identifie que les deux catégories suivantes de néologismes: les créations complètement nouvelles et les emprunts.

Quel que soit le mode de formation, un néologisme décrit une notion ou une réalité nouvelle. La terminologie vise à rassembler les données sur le lexique d'une langue de spécialité ainsi qu'à normaliser les termes (Sager 1990: 2). Les ressources terminologiques (glossaires, banques de termes) devraient fournir, dans la mesure du possible, une représentation de l'état actuel du domaine à l'étude, sur le plan aussi bien linguistique que notionnel. Mais souvent, tout au début de l'introduction d'une notion, la terminologie qui la décrit n'est pas fixe. Plutôt, d'après Sager (1990: 82), plusieurs formes coexistent. Donc, en premier lieu, il est important que le terminologue reconnaisse, recueille et enregistre toutes les formes. Au fil du temps, il peut suivre les tendances d'usage et il peut faire des recommandations en ce qui concerne la création ou la sélection de termes préférés. Du fait qu'elle représente un travail important, la procédure de détection des néologismes est longue, trop longue pour le terminologue occupé, qui a besoin d'un outil pour lui venir en aide.

3 *Cenit:* vue d'ensemble

Nous avons créé un outil qui s'appelle *Cenit* (*Corpus-based English Neologism Identifier Tool*). *Cenit* vise à identifier des néologismes dans la langue de spécialité. Dans les paragraphes suivants, nous décrivons les caractéristiques et le fonctionnement de ce système. Néanmoins, il faut dire tout d'abord que *Cenit* n'est encore qu'un prototype; donc il présente à l'heure actuelle quelques limites. Premièrement, nous avons choisi la langue anglaise comme point de départ, mais en principe, les techniques pourraient être adaptées au français. Deuxièmement, *Cenit* n'identifie que les néologismes de forme. Il ne peut pas détecter les néologismes de sens, créés, par exemple, par extension sémantique ou par changement de catégorie grammaticale. Le fait que *Cenit* ne traite que des unités lexicales qui consistent en un seul mot, qu'elles soient des créations complètement nouvelles ou des dérivations (créées par préfixation, suffixation, etc.) constitue une dernière limite du prototype. C'est-à-dire qu'actuellement, *Cenit* ne peut pas détecter les néologismes de forme syntagmatique, mais nous avons l'intention d'accroître la capacité du système à cet égard dans la phase suivante du projet.

Pour réaliser *Cenit*, nous avons choisi le langage de programmation *Perl* car celui-ci manipule très bien les chaînes de caractères. Le système fonctionne sous *Unix* avec l'accès par le *World Wide Web*. *Cenit* comprend deux programmes principaux: le programme pour créer les dictionnaires spécialisés, et le programme de recherche. Le système comprend un dictionnaire de base provenant du lexique Unix et le programme pour créer les dictionnaires spécialisés donne au terminologue la possibilité d'ajouter d'autres dictionnaires afin d'adapter *Cenit* pour traiter un domaine spécifique.

Cenit vise à traiter un fichier d'entrée (fournit par l'utilisateur) afin d'identifier les néologismes qu'il comporte. Le programme de recherche est le programme principal qui exécute les tâches suivantes. Il transforme le fichier d'entrée en liste de types (c.-à-d. une liste qui comprend chaque forme différente figurant dans le texte) qu'il fait passer par une série de filtres. Ces filtres examinent chaque forme et essayent de l'éliminer en la comparant avec les dictionnaires. Après avoir été filtrée, la liste est considérablement réduite. Les mots qui restent peuvent être des néologismes et chacun d'entre eux est exposé sur l'écran, souligné dans le contexte d'apparition, à charge pour le terminologue d'effectuer un tri.

3.1 Programme pour créer les dictionnaires spécialisés

Le dictionnaire de base est assez général et de plus, il ne contient que des formes canoniques (c.-à-d. les

formes non-dérivées et non-fléchies). Donc, nous anticipons que les terminologues aimeront construire des dictionnaires spécialisés en utilisant des corpus de textes présents à l'état naturel. À ce stade, *Cenit* prend comme données d'entrée un corpus de textes spécialisés fournit par l'utilisateur. Il transforme ce corpus en liste d'occurrences (*tokens*) et finalement en liste de types. Évidemment, plus le corpus est grand, plus le dictionnaire spécialisé sera complet et plus les résultats seront fiables. Notons que le programme n'ajoute pas au dictionnaire spécialisé les formes qui se trouvent déjà dans le dictionnaire de base. Chacun des dictionnaires comprend 52 fichiers triés: deux par lettre (aA-zZ) pour qu'on puisse accéder soit aux minuscules, soit aux majuscules.

3.2 Programme de recherche

Dans les paragraphes suivants, nous décrivons les filtres et les stades de traitement plus en détail. L'ordre dans lequel ce programme traite les données est extrêmement important. La figure 1 illustre l'ordre de l'application des filtres. Les crochets indiquent qu'un stade est facultatif.

3.2.1 Traitement du fichier d'entrée

La première étape dans le processus d'identification des néologismes est le traitement du fichier d'entrée. *Cenit* crée trois copies différentes du fichier d'entrée. Dans la première copie, tous les signes de ponctuation sont supprimés. Le système emploiera cette copie pour créer une liste d'occurrences (*tokens*) et finalement une liste de types. La deuxième copie ressemble à la première à la différence que les points (.) ne sont pas supprimés. *Cenit* emploiera la deuxième copie pour identifier les noms propres en déterminant si un nom avec une majuscule se trouve au début ou au milieu d'une phrase. La troisième copie est une copie exacte du fichier d'entrée d'origine où tout reste intact. Cette copie sera utilisée au dernier stade quand *Cenit* exposera les néologismes candidats dans leurs contextes d'apparition.

3.2.2 Exclusion des correspondances exactes

Le premier filtre prend comme données d'entrée la liste des types provenant de la première copie du fichier d'entrée. Chaque mot est comparé avec le dictionnaire de base et avec les dictionnaires spécialisés. Seuls les mots qui ne figurent pas dans les dictionnaires avancent à la prochaine étape.

3.2.3 Exclusion par l'analyse morphologique

Car le dictionnaire de base ne contient que les formes canoniques (par exemple, *manger*), l'utilisateur peut souhaiter exclure aussi les formes dérivées ou fléchies (par exemple, *mange*, *mangeons*, *mangé*, etc.). Cependant, la dérivation est l'un des modes les plus communs de formation des néologismes, c'est la raison pour laquelle, si l'on utilise ce filtre, on risque exclure les néologismes intéressants (par exemple, les néologismes hypothétiques *mangement* ou *mangeation*). Donc, cette étape n'est pas obligatoire. Nous supposons que la plupart des formes dérivées et fléchies «connues» se trouvent dans le dictionnaire spécialisé créés par le terminologue car la source de ce dictionnaire est un corpus de textes présents à l'état naturel. Par conséquent, nous imaginons que cette option ne sera employée que par les utilisateurs qui n'ont pas de dictionnaire spécialisé.

À ce stade, les mots sont déjà traités par le filtre des correspondances exactes, ils ne se trouvent donc pas ni dans le dictionnaire de base, ni dans le dictionnaire spécialisé. Si on lance le filtre d'analyse morphologique, chaque mot est traité pour en déterminer la racine et si la racine se trouve dans l'un des dictionnaires, *Cenit* supprime la forme dérivée ou fléchie.

3.2.4 Séparation entre les mots qui commencent par une majuscule et les mots en minuscules

Ce stade prend comme données d'entrée la liste du stade précédent. Les mots sont divisés en deux groupes: le premier groupe comprend les mots qui commencent par un caractère majuscule, et le deuxième comprend les mots qui ne commencent pas par une majuscule. Les trois stades suivants ne traitent que les mots du premier groupe.

3.2.5 Exclusion des correspondances exactes sans faire attention au type de caractère

Toutes les majuscules sont transformées temporairement en minuscules. Ensuite, chacun des mots est comparé à nouvelle reprise avec les dictionnaires. Cette seconde comparaison vise à supprimer les mots qui n'avaient de majuscule que parce qu'ils étaient placés en tête de phrase.

3.2.6 Exclusion des initialismes

La création d'un néologisme par l'acronymie est valable, surtout si le sigle se prononce comme un mot (par exemple, *laser*). Lorsqu'ils sont reconnus comme mots, ces acronymes s'écrivent souvent en minuscules. Néanmoins, plusieurs terminologues ne s'intéressent pas aux sigles qui ne comprennent que des initiales tout en majuscules et qui ne se prononcent pas comme un mot (par exemple, CPTIAQ). Ce type de sigle

correspond souvent au nom d'un produit, d'une société, etc. *Cenit* propose alors un stade facultatif qui vise à supprimer les initialismes tout en majuscules. Pour chacun des mots sur la liste, le système calcule le nombre total de caractères ainsi que le nombre de majuscules, et si les deux sont égaux, *Cenit* supprime le mot.

3.2.7 Exclusion des noms propres

Ce filtre se divise en deux parties. D'abord, *Cenit* compare le mot qui précède le néologisme candidat avec une «liste d'arrêts» (stoplist) qui comprend les mots comme Monsieur, Mme, Professeur, Dr, etc. Si le mot précédent se trouve sur la «liste d'arrêts», le néologisme candidat est supprimé. Pour ceux qui restent, ils procèdent à l'étape suivante, qui emploie la deuxième copie du fichier d'entrée (c.-à-d. la version qui ne contient aucune ponctuation sauf les points). D'abord, le système repère précisément l'emplacement dans le texte de chaque néologisme candidat. Si le candidat n'est pas précédé par un point (c.-à-d. s'il se trouve au milieu d'une phrase), il est probable que le candidat est un nom propre et, par conséquent, *Cenit* le supprime.

3.2.8 Regroupement des mots en majuscules et des mots en minuscules

Les néologismes candidats qui ont des caractères majuscules sont maintenant réunis avec les candidats en minuscules. Cette liste représente la somme de néologismes candidats qui se trouvent dans le fichier d'entrée.

3.2.9 Exposition des néologismes candidats et de leurs contextes d'apparition

Le dernier stade consiste à exposer chacun des néologismes candidats en contexte d'apparition. *Cenit* repère l'emplacement de chaque candidat dans la troisième copie du fichier d'entrée (c.-à-d. la copie exacte du fichier d'origine). Ensuite, le système identifie les limites de la phrase en cherchant les points et les caractères majuscules. Pour finir, *Cenit* fournit une liste de néologismes candidats en contexte d'apparition, à charge pour le terminologue d'effectuer un tri. (Noter que le terminologue peut regarder les résultats à la fin de chaque stade, s'il le désire). Ensuite, le terminologue peut ajouter au dictionnaire tout mot qui fait partie de la liste mais qui n'est pas néologisme.

3.3 Extension du prototype actuel

Le prototype actuel est trop limité pour répondre à tous les besoins des terminologues. À court terme, nous prévoyons de procéder à des extensions, comprenant la possibilité d'identifier des néologismes syntagmatiques et, peut-être, la possibilité de détecter les néologismes qui sont créés par changement de catégorie grammaticale. Pour identifier les syntagmes, nous avons l'intention d'étudier les bigrammes et les trigrammes dans les corpus. Pour détecter les changements de catégorie grammaticale, nous proposons d'examiner les collocations de termes.

4 Conclusion

Nous avons affirmé, au début de cet article, que les langues évoluent au rythme accéléré. Détecter manuellement les néologismes n'est plus pratique pour les terminologues, qui doivent souvent travailler dans des délais très serrés. Les ordinateurs se sont révélés utiles pour aider les terminologues à réaliser un travail important dont la détection des néologismes fait partie. C'est une époque bien excitante pour la terminologie et nous espérons que le travail décrit ici pourra servir de point de départ à des recherches plus approfondies dans ce domaine.

Sorcha Roche et Lynne Bowker,
School of Applied Language and Intercultural Studies,
Dublin City University,
Irlande.

Remerciements

Nous remercions Danièle Tort (DCU.LS) pour la correction des épreuves.

Bibliographie

Blackwell (S.), 1993: «From Dirty Data to Clean Language», dans Aarts (J.), de Haan (P.), et Oostdijk (N.), dir., *English Language Corpora: Design, Analysis and Exploitation*, Amsterdam, Rodopi, p. 97-105.

Bourigault (D.), Jacquemin (C.), L'Homme (M.-Cl.), dir., 1998: Computerm '98, *Proceedings of the First Workshop on Computational Terminology, COLING-ACL '98*, 15 août 1998, Université de Montréal, Canada.

Collier, (A.), 1993: «Issues of Large-Scale Collocational Analysis», dans Aarts (J.), de Haan (P.), et Oostdijk (N.), dir., *English Language Corpora: Design, Analysis and Exploitation*, Amsterdam, Rodopi, p. 289-298.

Cabré (M.T.) et de Yzaguirre (L.), 1995: «Stratégie pour la détection semi-automatique des néologismes de presse», dans *Traduction, terminologie, rédaction*, vol. 8 (2), p. 89-100.

Condamines (A.), 1995: «Analyse de textes spécialisés pour le recueil de données terminologiques», dans *Terminologies nouvelles*, n° 14, p. 35-42.

Dubuc (R.), 1985: *Manuel pratique de terminologie* (*2e édition*), Québec, Linguatech.

Janicijevic (T.) et Walker (D.), 1997: «NeoloSearch: Automatic detection of neologisms in French Internet documents», présenté au colloque ACH-ALLC'97, Kingston, Canada, (http://www.quics.queensu.ca/achallc97/papers/a009.html [date d'accès 03/09/98])

Kageura (K.) et Umino (B.), 1996: «Methods of Automatic Term Recognition: A Review», dans Terminology, vol. 3 (2), p. 259-289.

Le Meur (A.) et Depecker (L.), 1995: «Balnéo: un projet de réseau informatique pour la veille néologique», dans *Terminologies nouvelles*, n° 14, p. 48-53.

Miller (D.), Meyer (I.) et Michaud (D.), 1991: «Terminologie et analyse notionnelle assistée par ordinateur», dans *Actes du Colloque international sur les industries de la langue*, Tome II, Québec, Office de la langue française et STQ, p. 781-800.

Renouf (A.), 1993: «A word in time: First findings from the investigation of dynamic text», dans Aarts (J.), de Haan (P.), et Oostdijk (N.), dir., *English Language Corpora: Design, Analysis and Exploitation*, Amsterdam, Rodopi, p. 279-288.

Rey (A.), 1995: *Essays on Terminology* (traduit par J.C. Sager), Amsterdam, John Benjamins.

Rondeau (G.), 1984: *Introduction à la terminologie* (*2e édition*), Québec, Gaëtan Morin.

Sager (J.C.), 1990: *A Practical Course in Terminology Processing*, Amsterdam, John Benjamins.

Sites Internet

Aviator, University of Liverpool, Royaume-Uni: http://radar.rdues.liv.ac.uk/newwds.html [03/09/98]

Cenit, Dublin City University, Irlande: http://www.compapp.dcu.ie/Projects/1998/CL/sroche.cl4/ [03/09/98]

Cordon, Institut für deutsche Sprache, Allemagne: http://www.ids-mannheim.de/telri/seminar/jv-cordon.html [03/09/98]

Neolosearch, Queen's University, Canada: http://www.quics.queensu.ca/achallc97/papers/a009.html [03/09/98]

Obneb, Universitat Pompeu Fabra, Espagne: http://www.iula.upf.es/pjobneca.htm [03/09/98]

Figure 1: Les stades du programme de recherche CENIT

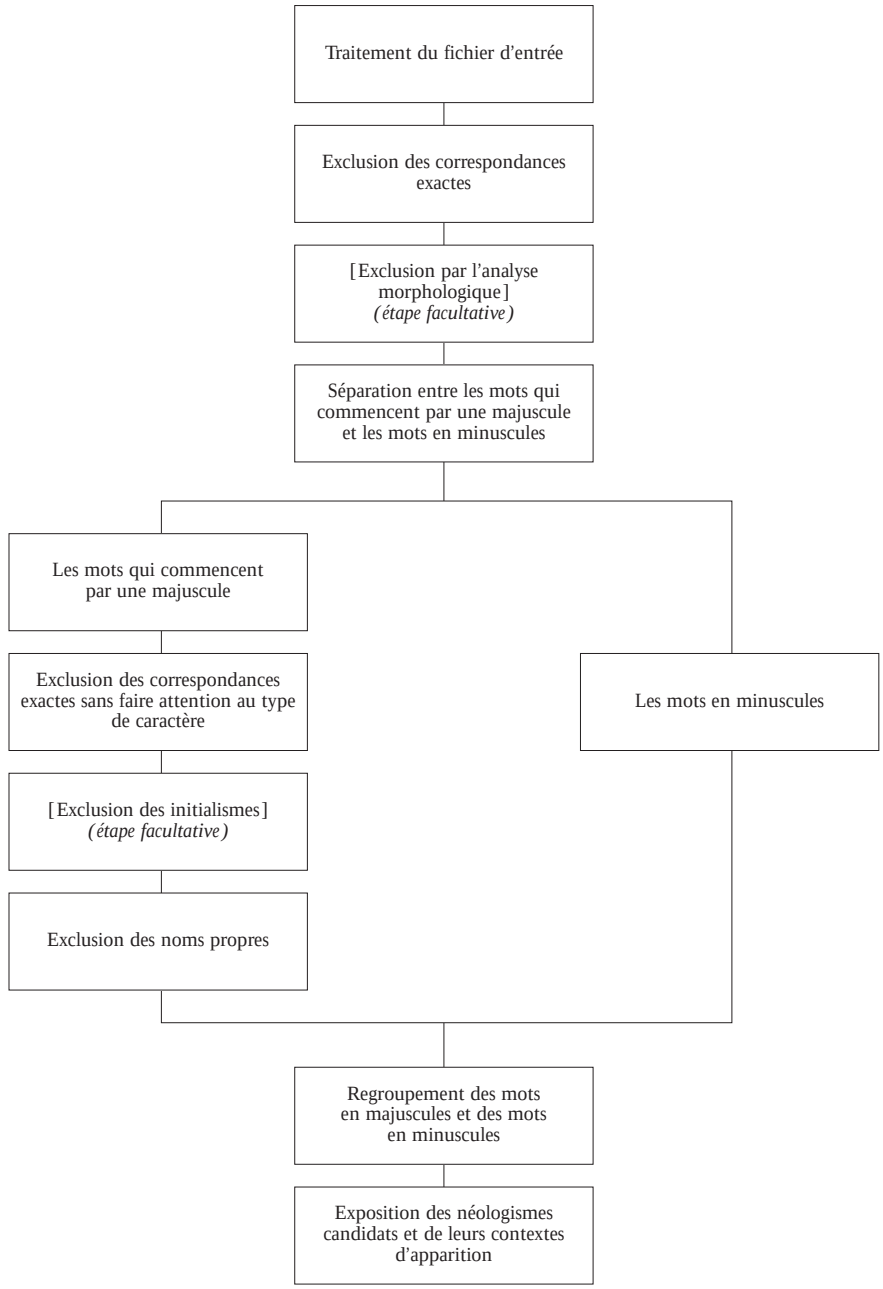

Figure 1: Les stades du programme de recherche *Cenit*