

La question épineuse de l'évaluation des outils informatiques dans le cadre de travaux terminologiques

ROSA CETRO

Université de Catania (S.D.S. di Ragusa)

INTRODUCTION

Bien que des pratiques terminographiques soient nées déjà à la Renaissance, il a fallu attendre le XX^e siècle pour qu'une théorie de la terminologie soit développée. Les nouvelles réalités créées par la société industrielle et post-industrielle se sont concrétisées dans le besoin de disposer de nouveaux termes pour les nommer et ont ainsi favorisé le développement de travaux en terminologie. En même temps, une autre discipline fondamentale pour les sociétés modernes a vu le jour au XX^e siècle : il s'agit de l'informatique, née pendant la Seconde guerre mondiale. A partir des années 1960, terminologie et informatique ont entrepris une collaboration qui est devenue de plus en plus étroite, au point qu'il n'est presque plus possible de parler de terminologie sans parler d'informatique.

Ainsi, de nos jours, la plupart des travaux en terminologie s'appuient sur des outils informatiques, conçus en vue des applications les plus variées (confection de dictionnaires, réalisation d'ontologies, extraction d'information, etc.). Généralement, les performances de ces outils font l'objet d'une évaluation, souvent en termes de *rappel* et *précision*.

Dans le cadre de notre thèse de doctorat, nous avons testé trois outils, différents entre eux tant pour les méthodes qu'ils exploitent que pour les types de tâches qu'ils accomplissent. Devant évaluer les tests menés avec ces logiciels sur

deux corpus comparables spécialisés, nous avons touché du doigt les difficultés de la tâche d'évaluation, que nous qualifierions d'épineuse, car il est plutôt réducteur de la limiter à une simple valeur numérique.

Après un rapide survol sur l'histoire des rapports entre terminologie et informatique, nous présenterons nos expériences sur ces trois logiciels et nous commenterons quelques aspects – pratiques et théoriques – liés à la question de l'évaluation.

1. DES BANQUES DE TERMINOLOGIE À LA NAISSANCE DE LA TERMINOLOGIE COMPUTATIONNELLE

Suivant une répartition commune à plusieurs chercheurs, parmi lesquels nous citons Drouin (2002), on reconnaît trois grandes étapes de la collaboration entre terminologie et informatique : une première étape qui a vu le développement des banques de terminologie (années 1960-1980), une deuxième étape marquée par l'apparition des outils de terminotique et la naissance des industries de la langue (années 1980) et une troisième étape (des années 1990 à nos jours) caractérisée par la naissance et le développement de la terminologie computationnelle.

Les banques de terminologie sont le fruit de l'avancement des techniques documentaires combiné aux progrès de l'informatique lourde. Les données terminologiques y sont stockées sous formes de fiches. Ce premier système de stockage est mis en place en Europe avec le projet DICAUTOM (1963), un grand dictionnaire sur support électronique créé par la CECA (Communauté Européenne du Charbon et de l'Acier). Au début, ces systèmes sont réalisés surtout dans le cadre de projets gouvernementaux et institutionnels. Mais ils attirent bientôt l'intérêt de la part des entreprises, pour les avantages qu'ils offrent à la gestion de l'information.

Les années 1980 sont cruciales pour l'intensification des rapports entre terminologie et informatique. Deux concepts-clé voient le jour, ceux de *terminotique* et d'*industries de la langue*.

Auger (1989 : 450) parle de la terminotique (ou terminologie automatique) comme d'une « nouvelle composante lourde de la terminographie moderne ». Plus récemment, dans le but d'opérer une comparaison avec la terminologie computationnelle – dont nous parlerons plus loin – L'Homme (2004 : 17) définit la terminotique comme « l'ensemble des activités liées à la description des termes dans lesquelles intervient une application informatique ». Le concept se diffuse dans la deuxième moitié de la décennie 1980 pour indiquer l'intégration des banques de terminologie dans des outils de bureautique, grâce à des supports – cédéroms, disquettes – qui en permettent la consultation depuis des micro-ordinateurs.

Pour ce qui concerne l'expression « industries de la langue », elle apparaît au singulier et entre guillemets dans un article que François Dégremont, rattaché à la Mission interministérielle de l'information scientifique et technique (MI-

DIST), publiée dans la revue *Brises* en 1984. L'auteur présente ce domaine comme riche en activités et applications, parmi lesquelles la terminologie revêt une importance capitale. Une définition plus précise des industries de la langue est donnée trois ans plus tard, au Sommet de la Francophonie de Québec (1987), en vue duquel le Comité québécois des industries de la langue (CQIL) affirme que

Les industries de la langue sont celles qui conçoivent, fabriquent et commercialisent des appareils et logiciels qui manipulent, interprètent, génèrent le langage humain, aussi bien sous sa forme écrite que sous sa forme parlée, en se fondant sur les travaux et les recherches des sciences du traitement de l'information et du langage (Rapport CQIL 1987 : 73)¹.

La terminotique revendique sa place dans ce panorama, en tant que champ de recherche de l'informatique linguistique. Les raisons de son développement sont à attribuer à la portabilité de l'informatique (micro-ordinateurs) et aux progrès de l'informatique à orientation textuelle. C'est dans ces années que le « poste du terminologue » voit le jour, ceci grâce aussi aux accords entre certaines entreprises et les universités, comme l'accord stipulé entre IBM et l'Université Laval (Québec) en 1985.

Pour Auger (1989) l'évolution de la terminologie est inévitablement liée à l'informatique, comme pour d'autres disciplines reposant sur le traitement de l'information écrite. Illustrant les différentes phases du travail terminographique, le linguiste-terminologue explique comment l'informatique pourrait en alléger et accélérer certaines tâches. Toutefois, encore à la fin de la décennie 1980, un empêchement majeur à l'automatisation (ou même la semi-automatisation) de la terminographie est constitué par le manque de textes spécialisés sous format électronique. Les terminographes créent des corpus électroniques en les saisissant à la main sur les ordinateurs ou bien en recourant à des systèmes de lecture optique pour la saisie automatique de caractères, qui ne résolvent le problème qu'en partie.

Si encore à la fin des années 1980 on déplore un manque de textes sous format électronique, à peine quelques années plus tard ce problème semble être résolu. Bourigault et Jacquemin (2000) identifient trois facteurs à la base de la prolifération de documents sous format électronique : l'internationalisation des échanges, la diffusion des outils de bureautique et le développement d'Internet. C'est dans les années 1990 que l'on assiste au passage de la terminotique à la terminologie computationnelle

[...] qui croise les apports de la terminologie, de la linguistique, du TAL, et de l'ingénierie des connaissances, c'est-à-dire des instruments, dispositifs et méthodes de recueil et d'utilisation des connaissances. Cette terminologie computationnelle allie plusieurs types de travaux (Habert 2005 : 78).

¹ Tant la citation de Dégremont que celle du CQIL sont tirées respectivement de Corbeil (1990 : 8) et de Corbeil (1990 : 18).

L'augmentation de la documentation au sein des institutions et des entreprises a pour conséquence de nouveaux besoins de gestion de l'information : les ressources terminologiques gagnent en importance et ne sont plus limitées aux dictionnaires spécialisés ou aux banques de termes. De nouveaux produits terminologiques sont créés, pour répondre aux nouvelles applications de la terminologie en entreprise. Dans ce contexte, l'acquisition de terminologie à partir de corpus devient une priorité. Ainsi, les premiers logiciels d'extraction de terminologie font leur apparition, suite aussi au regain d'intérêt pour les travaux en analyse statistique de la langue relevé au début des années 1990 (Bourigault, Jacquemin 2000 : 217).

2. CLASSIFICATION DES OUTILS INFORMATIQUES UTILISÉS EN TERMINOLOGIE

Deux caractéristiques distinguent les premiers travaux liés à la conception d'outils d'acquisition de terminologie : ils sont menés sur la langue française et en contexte industriel. Le premier outil visant exclusivement la construction de bases de données terminologiques est le progiciel *Termino* (réalisé par S. David et P. Plante en 1990), fruit d'une collaboration entre l'Université du Québec à Montréal et l'OQLF (Office Québécois de la Langue Française). Le fait que le français soit la langue sur laquelle se concentrent ces types de recherches s'explique essentiellement par deux raisons, l'une de nature plus proprement linguistique, l'autre de nature pour ainsi dire politique. D'un point de vue linguistique, à la différence de l'anglais, le français pose plus de problèmes dans le repérage automatique des unités terminologiques complexes, en raison de l'usage fréquent de prépositions et déterminants. Sous l'angle des politiques linguistiques, l'implémentation des outils et des ressources terminologiques en langue française se configure comme un moyen efficace pour contraster la concurrence de la langue anglaise.

Les tâches accomplies par les logiciels développés pour l'acquisition de terminologie à partir de corpus sont essentiellement trois : l'extraction terminologique, la structuration de terminologie et l'alignement de termes (dans le cadre de travaux bi- ou plurilingues). Sur le plan chronologique, les extracteurs de terminologie ont précédé les autres types d'outils. Il existe des outils qui accomplissent plusieurs tâches à la fois. L'analyse menée par ces outils peut se fonder sur des méthodes statistiques, des méthodes linguistiques (dites aussi symboliques) ou sur des méthodes hybrides, combinant les deux.

Les outils basés sur des méthodes statistiques trouvent leur origine dans les modèles mécaniques, utilisés en documentation à la fin des années 1980. Etant fondés sur des algorithmes, les outils statistiques n'exploitent pas de ressources lexicales, telles des grammaires ou des dictionnaires, parce qu'ils reconnaissent des chaînes de caractères et non pas des mots. Le critère principal sur lequel ils s'appuient est la fréquence : si une chaîne de caractères, c'est-à-dire un mot, apparaît souvent dans un corpus, il y a de fortes probabilités pour que celle-ci soit

identifiée comme candidat terme par le logiciel et donc extrait. Normalement, un seuil minimal de fréquence est établi par l'utilisateur du logiciel pour l'extraction des candidats termes. Par exemple, si on programme un seuil minimal de 3, le logiciel procèdera à l'extraction de candidats termes apparaissant au moins 3 fois dans le corpus. L'avantage principal des méthodes statistiques est leur bas coût. Ces méthodes sont particulièrement performantes sur les gros corpus.

En opposition aux approches statistiques, les approches linguistiques (ou symboliques) s'appuient sur des ressources exploitant des informations syntaxiques, lexicales ou morphologiques pour mener leurs analyses. Ces ressources peuvent être des dictionnaires électroniques ou des grammaires locales. Un des points forts des approches linguistiques concerne la qualité des descriptions fournies, qui sont généralement plus fines que celles des approches statistiques. Cet avantage est dû aussi à la variété des ressources lexicales exploitées dans l'analyse. En outre, les approches linguistiques permettent de traiter des corpus de petite taille.

De plus en plus d'outils font recours à des méthodes hybrides, combinant un filtrage statistique à une analyse de type linguistique. En général, les techniques linguistiques utilisées dans les approches hybrides ont recours à l'analyse syntaxique. L'ordre d'application des différents types de techniques varie selon les outils.

Nous renvoyons à Cetro (2011 : 51-53) pour une description plus détaillée des différents types de méthodes et pour la présentation de quelques outils.

3. L'ÉVALUATION DES PERFORMANCES DES OUTILS INFORMATIQUES

Les outils informatiques conçus pour des applications linguistiques font l'objet d'une évaluation de la part de la communauté des utilisateurs. Les mesures d'efficacité utilisées pour évaluer les performances de ces outils s'appellent *rappel* et *précision*. Le rappel représente le pourcentage de termes pertinents extraits par le logiciel en comparaison avec les termes pertinents du corpus, tandis que la précision équivaut au pourcentage des termes pertinents parmi la totalité des termes extraits. Ces deux mesures sont généralement inversement proportionnelles : à un taux élevé du rappel correspond normalement un taux bas de la précision. Il est donc plutôt rare d'atteindre des taux élevés de rappel et précision en même temps.

Complémentairement aux notions de rappel et précision, les notions de *silence* et de *bruit* indiquent respectivement le pourcentage de termes pertinents non repérés et le pourcentage des termes non pertinents à l'égard du corpus qui en revanche ont été extraits.

Pendant notre travail de thèse, nous avons eu l'occasion de tester quelques outils informatiques sur deux corpus comparables, en français et en italien, ayant trait à la médecine thermique. Ces deux corpus, qui ont été constitués à par-

tir de revues scientifiques et de dossiers de presse des stations thermales, sont de taille modérée – environ 178 000² mots pour le corpus français, environ 130 000 mots pour le corpus italien³ –, surtout si on les compare aux corpus utilisés de nos jours, qui peuvent atteindre plusieurs millions de mots. Ces tests des outils ont été suivis d’une évaluation des résultats fournis par les différents logiciels. Comme l’indique le titre de la présente contribution, nous avons qualifié d’épineuse cette étape d’évaluation, pour plusieurs raisons, différentes selon les outils testés.

Avant d’analyser ces expériences de logiciels, nous allons présenter les outils utilisés.

4. LES OUTILS TESTÉS : ANA, TERMOSTAT ET UNITEX

Les outils testés appartiennent aux trois catégories de logiciels citées plus haut : ANA (Enguehard 1993) est un outil statistique, TermoStat (Drouin 2002) s’appuie sur des méthodes hybrides, Unitex (Paumier 2002) est un logiciel basé sur des méthodes linguistiques. Ils se différencient également par le type de tâches accomplies : le premier est un extracteur de terminologie, le deuxième un dépouilleur de terminologie en ligne, le troisième est un logiciel d’analyse de textes, qui n’a pas été conçu expressément pour la terminologie.

4.1 TEST DU LOGICIEL ANA

ANA (Apprentissage Naturel Automatique), qui n’a jamais été distribué⁴, est un outil multilingue (limité aux langues non agglutinantes), basé sur deux méthodes algorithmiques. Le logiciel accepte en entrée des données brutes, qui n’ont pas été étiquetées préalablement. Le traitement des corpus dans ce logiciel se fait en deux modules, baptisés *familiarisation* et *découverte*. Suite au module de familiarisation, le corpus est nettoyé des signes de ponctuation et des signes diacritiques et les connaissances du corpus sont extraites et regroupées dans quatre listes. Cette étape sert à séparer les mots grammaticaux des candidats

2 Lorsque nous avons testé le logiciel ANA, en septembre 2010, nous avons fourni un corpus provisoire de textes français de taille légèrement inférieure au corpus français définitif utilisé dans notre thèse (il existe un écart d’environ 15 000 mots entre les deux corpus). Nous avons aussi demandé de traiter un texte court (1 500 mots environ), qui n’a pas pu être analysé en raison de sa petite taille.

3 Les deux corpus comparables contiennent aussi des articles qui nous ont été fournis par des médecins thermalistes français et italiens.

4 Malgré son indisponibilité sur le marché, il est toutefois possible de tester ANA en contactant par mail l’équipe de Chantal Enguehard, qui continue à travailler sur cet outil.

termes, qui font l'objet de la quatrième liste, appelée *bootstrap*⁵. Cette dernière est enrichie par induction dans le module de découverte : pendant cette deuxième phase, le logiciel recherche les cooccurrences d'évènements⁶ les plus récurrentes dans le corpus contenant au moins un des candidats termes contenus dans la liste *bootstrap*, qui est ainsi enrichie jusqu'au moment où le logiciel n'extrait plus de nouvelles cooccurrences.

Les résultats de l'extraction des termes faite par ANA sur le corpus français provisoire nous ont été fournis par courriel sous forme d'un tableau de texte. Ces résultats sont organisés par ordre alphabétique dans trois colonnes : dans une première, les candidats termes extraits ; dans une deuxième, le nombre d'occurrences ; dans la troisième, les segments de texte d'où le candidat terme a été extrait avec le nombre d'occurrences pour chaque segment :

| | | |
|-----------------|----|--|
| Cure de boisson | 31 | (cure de boisson, 30) (cures de boisson, 1) |
|-----------------|----|--|

Tableau 1 : exemple du tableau des résultats de l'extraction par ANA.

Les candidats termes extraits par ANA sont au nombre de 2085. Aucune différence n'est faite entre termes simples et complexes, tous les candidats termes sont rangés dans le même fichier. L'auteure nous a informée que le seuil minimal de fréquence établi pour l'extraction a été de 3 occurrences et que le nombre de termes présents dans le *bootstrap* était de 7, mais nous ne savons pas quels étaient ces termes. Pour le calcul du rappel, dans le fichier de départ les termes ont été isolés manuellement à l'aide de balises, comme dans l'exemple suivant :

<terme>phénomène de Raynaud</terme>.

Comme on peut l'imaginer, l'annotation manuelle de tous les termes du corpus soumis à l'analyse aurait requis beaucoup de temps. Nous avons donc mené le calcul du rappel sur une portion du corpus, réunissant des textes variés et dont la taille atteint environ 26 000 mots (16% du corpus). Il s'ensuit que le taux de rappel que nous reportons est un taux approximatif.

Les outils linguistiques de support à cette phase d'identification des termes présents dans la portion de corpus retenue pour le rappel ont été le *GDT (Grand Dictionnaire Terminologique)* et le *TLFi (Trésor de la Langue Française informatisé)*. Nous avons décidé de nous appuyer non seulement sur un dictionnaire terminologique tel le *GDT* mais de consulter aussi un ouvrage lexicographique général tel que le *TLFi* car de nombreux termes de domaines variés y sont répertoriés, sou-

5 Le terme *bootstrap* est souvent traduit en français par le terme *amorçage*. Toutefois, comme Enguehard emploie *bootstrap*, nous avons préféré garder ce terme.

6 Ces cooccurrences d'évènements dans ANA sont de trois types : 1) *expression* : il s'agit de la cooccurrence de deux termes, comme *cœur du réacteur* ; 2) *candidat* : c'est la cooccurrence d'un terme et d'un mot séparés par un mot de schéma, comme *cuve du barillet* ; 3) *expansion* : il s'agit dans ce dernier cas de la cooccurrence d'un terme et d'un mot, comme *structures internes*.

vent accompagnés d'une marque spécialisée. Ces mêmes outils nous ont servi lors du calcul de la précision.

1 504 termes ont été identifiés dans la portion de corpus choisie pour l'évaluation du rappel. Sur ces 1 504 termes, 527 figurent dans la liste des candidats termes fournie par ANA. Le taux de rappel approximatif est donc de 35%. En ce qui concerne la précision, sur les 2 085 candidats termes sortis par ANA nous en avons retenu 961, ce qui équivaut à un taux de précision de 46,09%.

Afin de procéder à la validation des candidats termes, nous avons dû établir des critères pour distinguer les résultats pertinents. Outre la pertinence sémantique, nous avons pris en considération la pertinence syntaxique, c'est-à-dire les limites du découpage en ce qui concerne les termes complexes. Pour le critère de pertinence sémantique, nous avons retenu toutes les séquences ayant un statut terminologique⁷ dans le corpus, c'est-à-dire que le choix n'a pas été limité aux techniques et aux moyens thermaux, mais a été élargi également à des termes de la médecine et d'autres domaines connexes au domaine thermal (chimie, pharmacologie).

Parmi les résultats de l'extraction terminologique opérée par ANA il y avait aussi bon nombre de candidats termes qui n'étaient pas du tout pertinents : par exemple, le logiciel a extrait des séries de chiffres (années, codes, etc.) ou des noms propres de stations thermales, uniquement en raison de leur fréquence dans le corpus.

L'évaluation des résultats fournis par ANA a été quelque peu complexe, en raison du fait que ce logiciel, étant dépourvu de concordancier⁸, ne fournit pas le contexte dans lequel un candidat terme apparaît. Surtout, elle a mis en lumière une conception théorique du terme qui ne nous trouve pas d'accord : le terme est ici considéré comme une unité dotée d'une valeur spécialisée définie a priori, alors que nous rejoignons la position de ces auteurs qui affirment que le terme acquiert cette valeur spécialisée lorsqu'il est employé dans un discours. Nous aurons toutefois l'occasion de discuter cet aspect plus loin.

4.2 TEST DU LOGICIEL TERMOSTAT

Contrairement à ANA, bien que développé initialement pour un projet d'entreprise, TermoStat est disponible en ligne et il est possible de le tester en créant son

⁷ Par *statut terminologique* on entend en terminologie le fait qu'une unité lexicale puisse être qualifiée de terme.

⁸ Les concordanciers sont des outils informatiques non spécifiques aux pratiques terminographiques mais qui offrent un large éventail de champ d'application, et dont le développement est allé de pair avec le regain d'intérêt pour les corpus en linguistique (dernier quart du XX^e siècle). Le résultat de l'application d'un concordancier à un texte produit une concordance, c'est-à-dire une liste des occurrences répondant à une recherche ciblée de la part de l'utilisateur.

propre espace personnel sur le site de référence⁹. Les langues prises en charge par le logiciel sont l'anglais, le français, l'espagnol, l'italien et le portugais. La détection des termes passe par des tests statistiques, menés à partir de la comparaison entre un corpus d'analyse (spécialisé) et un corpus de référence non technique (journalistique). Le logiciel liste les fréquences des unités lexicales des deux corpus : les unités dont la fréquence dans le corpus d'analyse est nettement supérieure que dans le corpus de référence seront identifiées comme pivots lexicaux spécialisés (PLS) et constituent le point de départ du processus d'acquisition terminologique (Drouin 2002 : 5).

Le processus d'acquisition des termes dans l'outil se déroule en trois étapes :

1) le prétraitement des données : le texte est segmenté en unités lexicales par un procédé mécanique ;

2) l'acquisition des pivots lexicaux spécialisés : pendant cette phase, le logiciel reconnaît dans le corpus les *spécificités positives*, les *spécificités négatives* et les *formes banales*¹⁰ ;

3) l'acquisition des termes : cette phase prévoit deux étapes, une première de recensement des candidats termes et une deuxième de filtrage de ces derniers sur la base de leur structure syntagmatique.

Pour le prétraitement du texte, TermoStat recourt à l'étiqueteur morphosyntaxique TreeTagger, qui procède à la désambiguïsation des mots susceptibles d'appartenir à plusieurs catégories grammaticales. Suite à la phase de prétraitement, chaque unité du texte se voit assigner une seule étiquette syntaxique. L'étiquetage morphosyntaxique permet à l'utilisateur de mener des recherches plus ciblées. En fait, depuis l'interface de TermoStat on peut choisir la catégorie grammaticale des termes à extraire (nom, verbe, adjectif ou adverbe), outre le type des unités terminologiques (simples, complexes ou les deux). Nous avons limité la recherche aux unités terminologiques nominales, tant simples que complexes. L'extraction a donné une liste de 3 522 candidats termes. Les critères appliqués pour la validation sont les mêmes que ceux utilisés dans l'expérience avec ANA. 1 769 candidats termes (désormais CT) ont été validés sur les 3 522 extraits, pour un taux de précision égal à 50,22%¹¹.

L'interface des résultats dans TermoStat permet d'accéder à cinq fenêtres différentes : Liste des termes, Nuage, Statistiques, Structuration et Bigrammes. Dans la première, les données sur chaque candidat terme sont organisées en cinq colonnes. Dans la première colonne (candidat de regroupement) sont listés les

9 http://olst.ling.umontreal.ca/~drouinp/termostat__web/index.php. Le logiciel peut uniquement être utilisé en ligne, il ne peut pas être téléchargé.

10 Les spécificités positives sont des formes nominales ou adjectivales qui sont très probablement des mots techniques, les spécificités négatives sont des formes qui présentent dans le corpus d'analyse une fréquence moindre par rapport au corpus de référence, alors que par formes banales on entend ces spécificités qui n'ont ni une valeur positive ni une valeur négative dans le corpus d'analyse.

11 Pour l'évaluation de TermoStat, nous nous sommes limitée au calcul de la précision.

candidats termes, suivis de leur nombre d'occurrences dans le texte (fréquence) et de leur score de spécificité (colonne score, spécificité). La quatrième colonne (variantes orthographiques) liste les variantes que le logiciel a repérées pour chaque candidat, alors que la dernière colonne (matrice) en décrit la structure syntaxique. La fenêtre Nuage affiche la liste des 100 termes dont le score de spécificité est le plus élevé dans le corpus sous forme de nuage. En ce qui concerne la fenêtre Statistiques, elle affiche des données numériques sur le texte analysé, notamment le nombre de candidats termes extraits et la répartition de ces candidats termes suivant les structures syntaxiques (dénommées *matrices* dans le programme), avec nombre exact et pourcentage dans le corpus traité. Passons maintenant à l'analyse de la fenêtre Structuration, qui comporte 3 colonnes. Les deux premières, Candidat de regroupement et Fréquence, sont les mêmes que pour la fenêtre Liste des termes, alors que la troisième, Terme inclus, liste les unités terminologiques complexes que le logiciel a extraites pour chaque candidat de regroupement¹². Pour ce module, les performances du logiciel sont directement proportionnelles à la taille du texte soumis à l'analyse, comme nous avons pu le constater en soumettant au logiciel un texte court (environ 1 500 mots). Le module Structuration a attiré notre attention, pour l'intérêt qu'il peut représenter en terminologie. Toutefois, ce qui nous frappe est la quantité réduite de termes complexes fournie pour deux termes simples très fréquents dans notre corpus, tels que *bain* et *douche*. Il suffit de penser que l'analyse des concordances dans le corpus a abouti à 67 termes complexes pour *bain* et 53 pour *douche*, alors que TermoStat a extrait 15 termes complexes pour *bain* et 9 pour *douche*. Si le but de la terminologie est l'élaboration d'un glossaire des soins thermaux qui soit le plus exhaustif possible, cet objectif ne peut pas être poursuivi en se basant uniquement sur les résultats fournis par l'extracteur.

On peut bien imaginer que, s'il est relativement aisé de s'apercevoir des termes passés sous silence dans des corpus de taille modérée, il n'en est pas ainsi pour des corpus de taille imposante. Certes, tout dépend de l'application visée, mais on ne saurait nier que, si l'extraction terminologique permet de gagner du temps, en même temps elle réduit la marge de liberté de l'utilisateur, qui doit se limiter à la liste de candidats termes fournie par le logiciel, à moins qu'il ne décide de revenir sur le corpus et procéder à une extraction terminologique manuelle.

4.3 EXPÉRIENCES SUR LE CORPUS AVEC LE LOGICIEL UNITEX

Développé par Sébastien Paumier (2002), Unitex est un logiciel qui réunit différents programmes pour le traitement de textes en langues naturelles sur la base de ressources lexicales. Plus précisément, il s'agit de ressources issues des travaux du lexique-grammaire sur la langue française – des dictionnaires

¹² Le candidat de regroupement peut être tant un CT simple qu'un CT complexe.

électroniques, des tables et des grammaires locales – qui, grâce au réseau RELEX, ont été étendus à d'autres langues. Le logiciel est téléchargeable sous une licence LGPL¹³ depuis le site du LIGM¹⁴. De nombreuses langues sont disponibles dans Unitex, parmi celles-ci il y a aussi des langues agglutinantes comme l'arabe ou isolantes comme le thaï.

Comme nous l'avons déjà dit plus haut, Unitex n'est pas un logiciel développé en vue d'applications terminologiques. Néanmoins, il présente un potentiel intéressant en terminologie de par ses fonctionnalités de recherche de contextes.

Lorsqu'on soumet un texte au logiciel, trois opérations sont exécutées pendant la phase de prétraitement : le comptage des formes du texte, l'étiquetage de ces formes, la segmentation du texte en phrases. Les résultats de ces opérations sont affichés dans trois fenêtres différentes. Ainsi, dans la première fenêtre (Token List) sont listées toutes les formes¹⁵ présentes dans le texte (signes diacritiques inclus) avec le nombre d'occurrences. Il est possible d'afficher la liste par fréquence (ordre décroissant) ou par ordre alphabétique. La deuxième fenêtre, Word Lists, est divisée en trois sous-fenêtres : une contenant les mots simples, une autre listant les formes composées (dans ces deux premiers cas, il s'agit de formes reconnues par les dictionnaires appliqués) et une dernière dans laquelle sont listées toutes les formes non reconnues par les dictionnaires. Les formes étiquetées se différencient des formes inconnues tout d'abord par l'utilisation de couleurs : bleu, rouge, vert et jaune. Elles sont suivies d'une série de codes morphosyntaxiques. En revanche, les formes inconnues des dictionnaires sont listées en noir et ne sont pas étiquetées. Nous avons remarqué que, dans le cas de textes spécialisés, l'observation des unités listées dans cette fenêtre était une première étape pour le repérage de termes.

Unitex n'accomplit ni d'extraction terminologique, ni de structuration de terminologie, mais il est équipé d'un concordancier de bonne qualité, qui permet de mener des recherches ciblées. On applique ce concordancier par le biais du menu Locate Pattern. La recherche peut être menée de façons différentes. On peut rechercher, par exemple :

- une unité lexicale donnée,
- toutes les occurrences d'une catégorie grammaticale.

Il est aussi possible de mener d'autres types de recherche, comme la recherche par filtre morphologique, qui peut se révéler intéressante pour la détection des composés savants (à titre d'exemple : on veut chercher dans le corpus toutes les unités lexicales commençant par le radical *hémo-* ou les mots contenant le suffixe *-icide*).

13 Lesser General Public License for Linguistic Resources.

14 <http://infolingu.univ-mlv.fr/>

15 *Token* ne correspond pas à *mot* : il s'agit de n'importe quel caractère du texte, même un espace est un *token*.

A partir du menu *Locate Pattern* on peut aussi mener une recherche par l'application de grammaires locales, qui sont des transducteurs représentables sous forme de graphes d'automates finis. Concrètement, les grammaires locales sont des motifs lexicaux ou syntaxiques que l'utilisateur veut rechercher dans un corpus textuel sous forme de boîtes reliées entre elles. Par exemple, on veut rechercher toutes les occurrences des mots *bain*, *douche*, *étuve* et *massage* suivis d'un adjectif, ou bien on veut rechercher toutes les séquences Nom de Nom présentes dans le corpus. Nous nous sommes servis des grammaires locales surtout pour le repérage des termes composés de nos deux corpus. Dans ce cas, comment procéder à l'évaluation des performances d'Unitex ? Peut-on appliquer les mêmes critères et les mêmes mesures d'efficacité utilisés pour l'évaluation des deux autres logiciels ? Il nous semble plutôt difficile, car une évaluation sur la couverture des grammaires locales serait plus une évaluation sur l'utilisateur qui les a réalisées que sur les performances du logiciel. De nouveau, la question de l'évaluation ne semble pas facile à aborder.

5. DES EXPÉRIENCES INFORMATIQUES À UNE RÉFLEXION THÉORIQUE SUR LE TERME

A la fin du paragraphe 4.1., nous avons affirmé que l'expérience avec le logiciel ANA a mis en lumière une conception théorique du terme que nous ne partageons pas. Il nous semble, en effet, qu'une telle présentation de l'extraction terminologique – une liste de mots donnés sans les contextes linguistiques dans lesquels ils apparaissent – reflète assez fidèlement la vision wüsterienne du terme, telle qu'elle est présentée dans la Théorie Générale de la Terminologie (TGT). D'après Wüster, les concepts, et non pas les termes, sont l'objet principal d'étude de la terminologie. Chaque concept se voit assigner un terme qui le désigne. La relation qui lie un terme à un concept est une relation de parfaite biunivocité. Dans ce modèle théorique, les termes se voient réduits à de simples étiquettes : ils n'ont pas le statut de signes linguistiques à part entière, leur fonction communicative et leur dimension syntaxique n'étant pas prises en considération.

Lors de l'évaluation des résultats du logiciel ANA, nous nous sommes souvent heurtée à la difficulté d'attribuer un statut terminologique à un candidat extrait. Si cette opération peut être aisée dans le cas de certains secteurs de la connaissance – par exemple, dans le cas des sciences exactes comme les mathématiques ou la chimie –, il n'en est pas de même dans de nombreux autres cas, en raison de la libre circulation des connaissances – et des termes qui les désignent – dans différents champs du savoir.

Il est intéressant de remarquer que, parallèlement au développement de la terminologie computationnelle, on a assisté à un véritable renouvellement théorique en terminologie. La TGT, demeurée pendant longtemps le modèle théorique de référence pour quiconque travaillait en terminologie, a fait l'objet d'une forte remise en discussion de la part des linguistes à partir du début des années

1990. Plusieurs modèles théoriques alternatifs ont été proposés : pour le monde francophone, nous citons la Socioterminologie de l'École de Rouen (représentée par Yves Gambier et François Gaudin) et la terminologie textuelle de Slodzian et Bourigault. Ce dernier modèle théorique est né dans le cadre de l'intelligence artificielle et a été présenté au Troisième Congrès du groupe TIA (Terminologie et Intelligence Artificielle), en 1999. En réalité, déjà en 1995, dans une communication au premier congrès en Terminologie et Intelligence Artificielle, Monique Slodzian (1995 : 11) illustre la nécessité de revisiter la doctrine terminologique sur la base des résultats de la lexicographie spécialisée, qu'elle juge « à maints égards insatisfaisants ». D'après l'auteure (*ibidem*), les insuffisances de ces produits découlent en large partie de l'approche théorique qui sous-tend leur réalisation :

Les difficultés tiennent en partie au fait que les milieux professionnels et institutionnels, auxquels sont principalement destinés les travaux de lexicographie spécialisée, ont une vision mécanique du couplage concept/mot et ne prennent pas en compte la complexité des phénomènes langagiers.

L'approche wüstérienne se présente comme une représentation taxinomique des connaissances qui, d'après Slodzian, est décidément dépassée par les recherches des sciences cognitives. Le « tout-paradigmatique » prôné par la théorie traditionnelle doit céder la place à un nouveau modèle théorique, hybride, intégrant le plan syntagmatique, donc l'étude du fonctionnement des termes dans les textes (Slodzian 1995 : 17). Les pistes ouvertes par la linguistique de corpus en terminologie sont incompatibles avec un modèle théorique, tel le modèle de la TGT, qui ignore la syntaxe du lexique.

Slodzian rebondit sur cette exigence d'un renouvellement théorique en terminologie dans l'intervention au Congrès de 1999 en collaboration avec D. Bourigault. Tout d'abord, les auteurs font un état des lieux des applications de la terminologie, dont le champ d'activité s'est décidément élargi suite à la diffusion des outils de bureautique et au développement d'Internet, entraînant la prolifération de documents sous format électronique. L'extension des applications de la terminologie, qui ne sont plus limitées aux glossaires spécialisés ou aux banques de données et qui jouent sur de plus grosses quantités d'informations, fait ressortir des problématiques méconnues des pratiques traditionnelles. La grande variabilité des applications de la terminologie va de pair avec le constat que les terminologies aussi sont variables. Ce qui revient à dire qu'il n'est plus possible de supposer l'existence d'une seule terminologie relative à un domaine d'activité, mais que le choix des unités à décrire dans une terminologie dépend de l'application visée. Le constat de la variabilité des terminologies remet en cause quelques piliers de la théorie wüstérienne : l'universalité des terminologies, le principe de biunivocité entre un concept et un domaine, le rôle de l'expert.

L'ensemble de ces constats empiriques entraîne des changements en profondeur de la pratique terminologique : l'activité de construction d'une terminologie est essen-

tiellement une tâche d'analyse de corpus textuels. Ils appellent du même coup à un renouvellement théorique de la terminologie : c'est dans le cadre d'une linguistique textuelle que doivent être posées les bases théoriques de la terminologie.

L'expérience montre que l'hypothèse selon laquelle l'expert d'un domaine serait le dépositaire d'un système conceptuel qu'il suffirait de mettre au jour est non productive. La tâche d'analyse terminologique vise alors avant tout la construction d'une description des structures lexicales à l'œuvre dans un corpus textuel à partir d'une analyse réglée de ce corpus (Bourigault, Slodzian 1999 : 30)

Comme on peut le constater par la citation ci-dessus, l'informatique a mis en évidence certaines limites de la théorie conceptuelle, en particulier la difficulté de traiter les termes sans tenir compte de leur dimension textuelle.

CONCLUSION

Dans cet article, nous avons présenté la problématique de l'évaluation de trois outils informatiques pour des applications terminologiques. D'une façon générale, nous avons constaté qu'il est plutôt réducteur de limiter les performances d'un logiciel à une valeur numérique. Suivant le logiciel, de plus, nous avons défini cette tâche d'épineuse pour différents aspects.

Dans le cas du premier logiciel, ANA, nous nous sommes confrontée à la difficulté d'établir le statut terminologique d'une unité lexicale extraite sans pouvoir disposer du contexte linguistique dans lequel elle s'insère. Cela reflète une conception du terme proche de celle adoptée dans l'optique conceptuelle, modèle théorique classique en terminologie, qui ne prend pas en considération la dimension linguistique des unités terminologiques.

L'évaluation du deuxième logiciel, TermoStat, a mis en lumière d'autres limites, comme par exemple la difficulté à cerner le taux de silence et aussi la pertinence partielle de certains candidats termes extraits (incomplets ou extraits uniquement sous leur variante la plus fréquente).

Pour ce qui concerne le dernier logiciel, Unitex, nous avons constaté qu'il était plutôt difficile d'adopter les mêmes paramètres employés dans l'évaluation des deux autres logiciels, pour plusieurs raisons : 1) le logiciel n'a pas été conçu expressément pour la terminologie, donc 2) il n'accomplit pas la tâche d'extraction terminologique, 3) l'évaluation des performances de l'application de grammaires locales serait plus une évaluation sur l'utilisateur qu'une évaluation sur le logiciel.

Si l'évaluation des logiciels est une tâche courante pour les informaticiens, elle peut s'avérer épineuse pour les linguistes, d'autant plus que, souvent, il est nécessaire d'utiliser plusieurs logiciels pour obtenir des résultats satisfaisants. Ces expériences sur les logiciels ont mis en évidence, par ailleurs, que, malgré cela, certaines tâches ne sont pas automatisables et que donc – heureusement ! – le dernier mot est toujours l'affaire du linguiste-terminologue ou de l'expert.

RÉFÉRENCES BIBLIOGRAPHIQUES

- Auger Pierre (1989), « La terminotique et les industries de la langue », *Meta*, 34/3, pp. 450-456.
- Bourigault Didier, Jacquemin Christian (2000), « Construction de ressources terminologiques », in Pierrel Jean-Marie (dir.), *Ingénierie des langues*, Paris, Hermès, pp. 215-233.
- Bourigault Didier, Slodzian Monique (1999), « Pour une terminologie textuelle », *Terminologies nouvelles*, 21, pp. 10-14.
- Cetro Rosa (2011), « Outils de traitement des langues et corpus spécialisés : l'exemple d'Unitex », in Dufiet Jean-Paul, Modena Silvia, Attruia Francesco, Cetro Rosa (éds.), *Cahiers de recherche de l'École Doctorale en Linguistique Française*, Milano, Lampi di Stampa, pp. 49-63.
- Corbeil Jean-Claude (1990), *Les industries de la langue : un domaine à la recherche de lui-même*, Québec, Gouvernement du Québec.
- Drouin Patrick (2002), *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*, thèse de doctorat, Université de Montréal.
- Enguehard Chantal (1993), « Acquisition de terminologie à partir de gros corpus », *Informatique et Langue Naturelle*, ILN'93, Nantes, pp. 373-384.
- Habert Benoît (2005), *Instruments et ressources électroniques pour le français*, Paris, Ophrys.
- L'Homme Marie-Claude (2004), *La terminologie : principes et pratiques*, Montréal, Les Presses de l'Université de Montréal.
- Paumier Sébastien, (2011), *Unitex 3.0. User manual*, Université Paris-Est Marne-la-Vallée, <http://infolingu.univ-mlv.fr/>.
- Slodzian Monique (1995), « Comment revisiter la doctrine terminologique aujourd'hui? », *La Banque des mots : Terminologie et Intelligence Artificielle*, numéro spécial 7, pp. 11-18.