Traduire
Revue française de la traduction

228 | 2013
Technique et pragmatisme

L'extraction terminologique à partir d'un corpus de textes techniques

Étude de cas appliquée au domaine de la sécurité informatique

Vasilica Le Floch

Édition électronique
URL : http://journals.openedition.org/traduire/537
DOI : 10.4000/traduire.537
ISSN : 2272-9992

Éditeur
Société française des traducteurs

Édition imprimée
Date de publication : 1 juin 2013
Pagination : 81-91
ISSN : 0395-773X

Référence électronique
Vasilica Le Floch, « L'extraction terminologique à partir d'un corpus de textes techniques », *Traduire* [En ligne], 228 | 2013, mis en ligne le 01 juin 2015, consulté le 10 décembre 2020. URL : http://journals.openedition.org/traduire/537 ; DOI : https://doi.org/10.4000/traduire.537

L'extraction terminologique à partir d'un corpus de textes techniques

Étude de cas appliquée au domaine de la sécurité informatique

Vasilica Le Floch

Le présent article propose une méthode d'extraction terminologique à partir de corpus bilingues comparables. Dans ce cas précis, il s'agit d'articles de spécialité du domaine de la sécurité informatique, articles écrits en langue anglaise et en langue française. Cette méthode fait appel à des outils disponibles sur internet et chaque étape est décrite afin d'offrir au lecteur une vision globale des procédés suivis. La méthode décrite peut être transposée à d'autres domaines et appliquée à d'autres corpus de spécialité, voire à une littérature de spécialité qu'un traducteur peut avoir à traiter dans son travail auprès d'organisations ou d'entreprises. Sans aucune prétention d'exhaustivité, l'article met en lumière une méthode rapide, basée sur des outils gratuits, donc disponibles pour tous.

Un nombre grandissant de textes de spécialité au format électronique sont disponibles sur internet. Le traducteur a ainsi facilement accès à des textes d'un domaine donné et cette base documentaire se développe de manière constante, étant donné la généralisation de la communication sur internet. La principale question qui se pose est de savoir comment exploiter ces ressources textuelles accessibles à tous, comment tirer profit de ces bases de données linguistiques et comment extraire des listes terminologiques susceptibles d'être utilisées par les traducteurs techniques travaillant dans un domaine donné.

Élaboration des corpus

Le point de départ de toute recherche terminologique d'un domaine de spécialité est la constitution d'une base de documents, ou d'un corpus. Idéalement, ce corpus sera constitué de documents techniques, propriété d'une entreprise cliente qui sollicite les services d'un traducteur technique. En l'absence d'une documentation fournie par la société cliente, et afin de constituer un glossaire minimal et de préparer le travail de traduction proprement dit, le traducteur peut se donner les moyens de constituer une base documentaire à partir des textes disponibles sur la toile. Bien que la première méthode présente un degré de spécialisation

plus important, les textes du corpus étant en totale adéquation avec les textes que le traducteur aura à traduire, deux éléments essentiels sont à noter : les sociétés clientes ne fournissent pas toujours une documentation de référence et le respect des exigences du droit d'auteur impose un certain nombre de précautions quant à l'exploitation des documents et à l'utilisation des termes et des glossaires qui en seront extraits.

Compte tenu des aspects énoncés précédemment, les corpus exploités pour les besoins de cette publication ont été constitués d'articles disponibles en ligne, sur *Wikipedia*. Afin de constituer des corpus spécialisés en langue française et en langue anglaise, cinq articles traitant du domaine de la sécurité informatique ont été retenus. Il faut noter que ces articles sont disponibles en français ainsi qu'en anglais, ce qui garantit une certaine homogénéité, autant en termes de taille que de contenu. Les textes retenus sont les suivants : Sécurité des systèmes d'information, Insécurité du système d'information, Système d'exploitation basé sur la sécurité, Vulnérabilité (informatique), Exploit (informatique). Leurs équivalents anglais, disponibles à partir de http://en.wikipedia.org/wiki/Main_Page, sont : *Computer security, Computer insecurity, Security-focused operating system, Vulnerability (computing), Exploit (computer security)*. Le corpus français contient 10 324 mots et l'anglais est légèrement plus important, avec un total de 11 289 mots. L'objectif est ici de composer deux corpus équivalents et le degré de spécialisation des textes est le critère qui a guidé notre sélection.

Deux corpus de référence, un pour le français et un pour l'anglais, ont également été constitués, chacun contenant environ 50 000 mots. Pour alimenter ces corpus, des articles d'ordre général comme Informatique, Communication, Histoire, Nature, Géographie ont été récupérés sur *Wikipedia*. Il faut noter au passage que ces corpus pourront être utilisés pour d'autres recherches terminologiques ultérieures. L'idée de base de cette approche est de pouvoir comparer le corpus spécialisé avec un corpus plus large et sans spécialisation particulière, afin de pouvoir identifier les termes qui n'apparaissent que dans le corpus de spécialité. Cette méthode est décrite en détail par Laroche et al. dans leur article « Étude de l'influence de la taille du corpus de référence sur l'extraction terminologique automatique contrastive » :

> *L'ETAC (extraction terminologique automatique contrastive) repose sur l'hypothèse selon laquelle les termes spécifiques au domaine ont des fréquences significativement plus élevées dans le CA (corpus d'analyse) que les fréquences attendues selon le CR (corpus de référence).* (p. 2)

Outils

La présente méthode d'extraction terminologique s'est fixé comme objectif d'identifier des outils simples et rapides, à la portée des traducteurs, si possible en évitant tout investissement financier. De nombreux logiciels d'aide à la traduction proposent des outils d'extraction termi-

nologique, qui peuvent parfois s'ajouter, comme option payante, au logiciel acquis initialement par le traducteur. Toutefois, il est possible de créer un premier glossaire en utilisant des logiciels gratuits et libres de droits, disponibles sur internet.

Le postulat de départ est le suivant : tout glossaire est le résultat du travail d'extraction de termes à partir d'un corpus donné. L'outil d'extraction le plus simple et le plus facile à manier est le concordancier. Il s'agit d'un logiciel qui, à partir d'un corpus donné, permet de faire le tri des mots, d'extraire les contextes pour chacun des mots et de générer des listes de fréquences ainsi que des index, pour ne citer que quelques fonctionnalités de base. Les listes de fréquences et les index sont des outils intéressants pour le traducteur, car ils permettent d'avoir rapidement une vue d'ensemble du vocabulaire d'un texte et de conduire, par la suite, des recherches terminologiques. Les listes de fréquences des mots d'un corpus constituent un indicateur précieux qui permet une première extraction grossière de candidats termes (voir à ce propos Thierry Hamon, 2011, Marie-Claude L'Homme, 2000).

Partant de ces considérations, nous avons testé plusieurs logiciels et avons retenu le concordancier *AntConc* version 3.2.4w, que son concepteur, Laurence Anthony, décrit ainsi : *« AntConc is a freeware, multiplatform tool for carrying out corpus linguistics research and data driven learning. »* (voir la documentation technique du concordancier, p. 1). Cet outil est distribué gratuitement et dispose de toutes les fonctionnalités de base d'un concordancier : concordance, liste de fréquences, index des mots présents dans le corpus, collocations (analyse du contexte d'un mot donné et identification des mots qui s'y trouvent de manière récurrente), syntagmes ou groupes de mots et mots clés. *AntConc* permet également de travailler sur plusieurs corpus ou textes et d'extraire les mots clés en comparant plusieurs textes ou bien en comparant un corpus d'analyse avec un corpus de référence que l'utilisateur aura constitué en amont. *AntConc* peut être téléchargé librement et l'installation ne prend que quelques minutes.

Traitement des corpus : listes de fréquences, concordances et tris contextuels

Une fois le logiciel d'*AntConc* installé, il est prêt pour l'utilisation, aucune authentification n'est requise. L'importation des corpus à analyser se fait de manière simple ; le format de fichier accepté est .txt. Une opération de nettoyage des textes s'impose lors de la préparation des fichiers .txt ; elle consiste à éliminer les signes qui ne sont pas pertinents pour l'analyse, tels que les chiffres, les lettres orphelines destinées au classement des listes, les étiquettes utilisées à l'intérieur des articles *Wikipedia* pour renvoyer à une autre page ou pour signaler une modification possible ou nécessaire d'un paragraphe. Une fois ce rapide nettoyage terminé, les corpus sont prêts à l'emploi et ils peuvent être importés dans *AntConc*. Le logiciel réunit sept

boutons sur une barre de tâches unique, ce qui facilite la navigation entre les différents tests appliqués au corpus. Pour les besoins de cette analyse, nous avons retenu les fonctions *Concordance, Collocates, Wordlist et Keyword List.*

L'analyse du corpus de langue française nous permet d'accéder rapidement à la liste des mots contenus dans le corpus, comme le montre la capture d'écran ci-dessous :

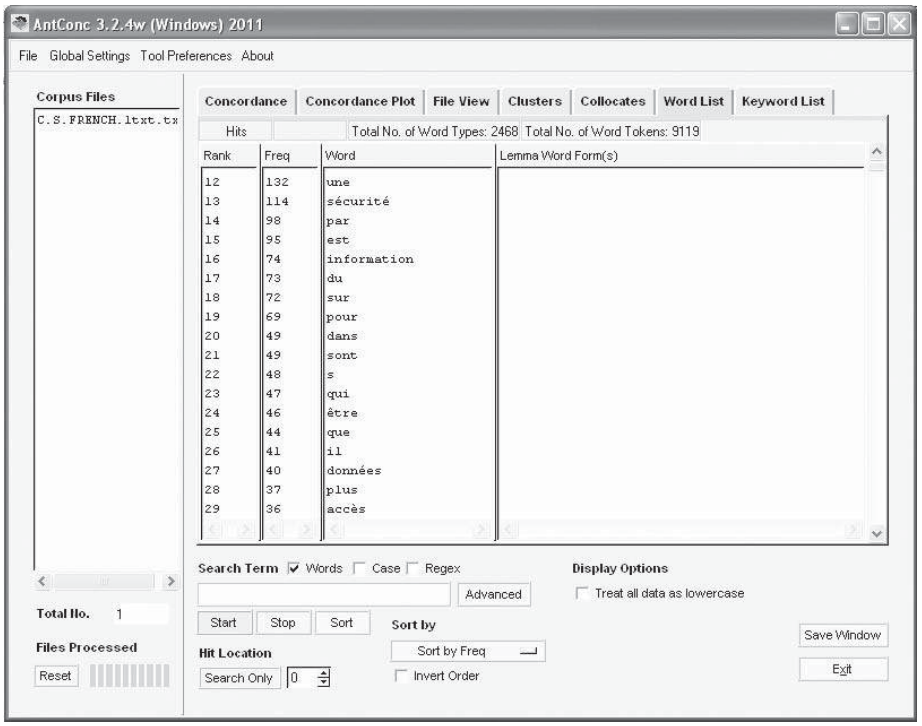

Figure 1 : Liste de mots dans le corpus français

Pour chaque mot de la liste, le logiciel affiche son rang (par ordre de fréquence) dans la colonne de gauche et sa fréquence dans la colonne du milieu. Le corpus français sur lequel nous avons travaillé contient 2 468 formes différentes. La classification des mots par ordre de fréquence nous offre un aperçu rapide des mots les plus fréquents utilisés dans le texte : si on élimine les mots grammaticaux (de, le, la, un, une, des, etc.), qui présentent, naturellement, les fréquences les plus élevées, la liste des mots à haute fréquence se présente comme suit :

sécurité 114	utilisateur 27
information 74	informatique 24
données 40	logiciels 24
accès 36	réseau 23
risques 28	vulnérabilités 23

Le corpus anglais soumis au même type d'analyse fournit des résultats légèrement supérieurs en ce qui concerne le nombre de formes différentes : il y a 2 534 formes dans le corpus, ce qui s'explique par la taille de celui-ci, supérieure à celle du corpus français. La classification des mots par ordre de fréquence, sans prendre en compte les mots grammaticaux (*the, to, of, and*, etc.), présente des fréquences élevées pour les termes suivants :

security 141	*access* 49
system 110	*software* 48
computer 79	*information* 46
systems 78	*vulnerability* 40
operating 52	*network* 34

Dans ces listes des dix mots les plus fréquents, on constate un taux de correspondance de 60 % entre les deux langues : les mots sécurité, système, information, accès, logiciels, réseau trouvent leurs équivalents dans le tableau des formes en anglais. Ces listes de fréquences permettent une première évaluation globale du vocabulaire des textes et une identification rapide de candidats termes pour l'élaboration d'un glossaire. Dans les deux listes, on remarque une quasi-totalité de substantifs, à une exception près, *operating*. Le logiciel *AntConc* permet d'accéder rapidement à la concordance d'un mot, simplement en cliquant sur celui-ci. Pour *operating*, la concordance dévoile une collocation récurrente : *operating system*. La fonction de recherche *Collocates AntConc* nous indique, en effet, que dans 48 des 52 occurrences d'*operating*, ce terme est suivi de *system*. Cette fonction permet donc de confirmer rapidement que nous sommes bien en présence d'un syntagme à inclure dans le glossaire. Il suffit à présent d'accéder à la concordance du mot « système » dans le corpus français pour voir les emplois de ce terme et les mots qui l'accompagnent, et identifier ainsi le syntagme « système d'exploitation ».

De la même manière, une recherche rapide au niveau du contexte du mot « logiciel » par le biais des concordances propose des syntagmes ou groupes de mots intéressants pour la recherche terminologique : « logiciel malveillant », « logiciel espion », « cassage de logiciel », et fournit également des équivalents anglais (tels qu'ils apparaissent dans les articles *Wikipedia* en français) à côté des mots français. On récupère ainsi *spyware* et *cracking* pour les deux derniers syntagmes et une vérification au niveau de la concordance du mot *software* dans le corpus anglais nous permet d'identifier le groupe *malicious software* comme syntagme équivalent de

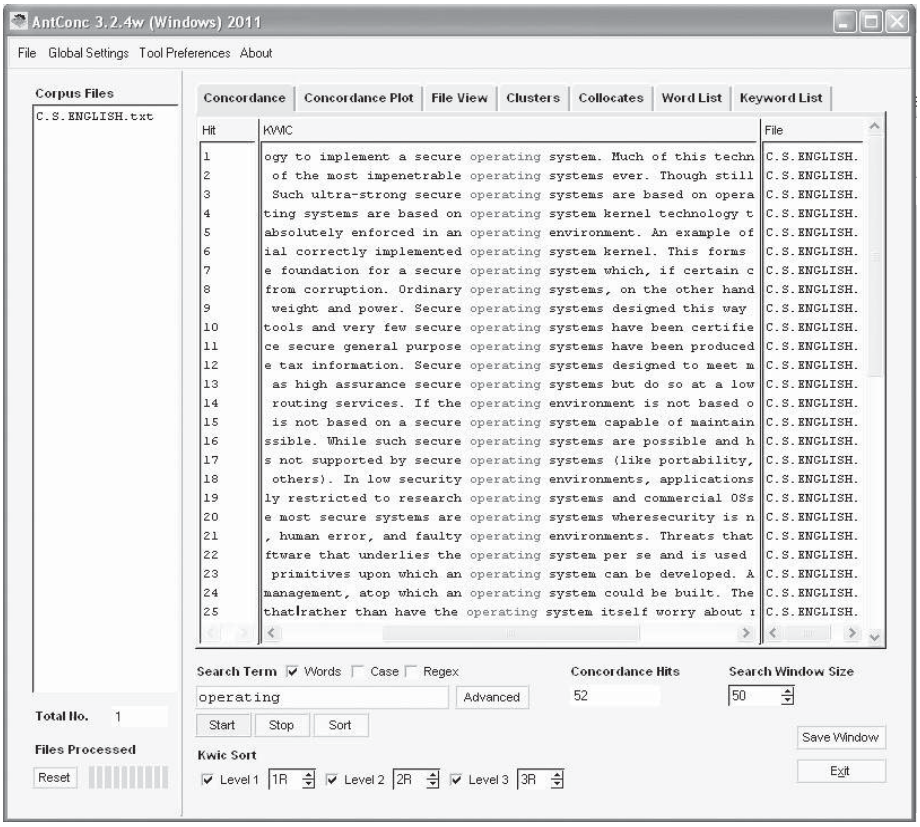

Figure 2 : Concordance du mot operating

« logiciel malveillant ». Ces premiers exemples, relativement simples, laissent penser que les concordances facilitent les recherches et les tris contextuels et permettent l'identification et l'extraction de termes de spécialité, avec un degré de précision satisfaisant. En continuant les tris et les sélections à partir des listes de fréquences de la manière décrite ici, le traducteur peut créer rapidement une liste de termes de spécialité du domaine de la sécurité informatique.

Extraction et analyse de termes : les mots clés

En complément de cette méthode simple d'extraction terminologique à partir des candidats termes des listes de fréquences, nous avons également testé la fonction *Keyword List AntConc*. Cette fonction est décrite par le créateur d'*AntConc* de la manière suivante :

> *This tool shows which words are unusually frequent (or infrequent) in the corpus in comparison with the words in a reference corpus. This allows you to identify characteristic words in the corpus, for example, as part of a genre or ESP (English for Specific Purposes) study.* (voir le fichier de documentation technique, p. 5)

Cette fonction nécessite donc l'intégration dans le logiciel d'un corpus de référence. Nous utilisons ici les corpus de référence constitués préalablement, tels qu'ils sont décrits dans la partie « Élaboration des corpus » du présent article. La liste des mots clés est donc extraite de manière automatique, sur la base de comparaisons avec le corpus de référence. *AntConc* propose deux méthodes d'extraction : la vraisemblance logarithmique *(log-likelihood)* et le chi carré *(chi square)*. Pour nos tests, nous avons utilisé la vraisemblance logarithmique, étant donné que le concepteur *AntConc* conseille l'utilisation de cette méthode. Le choix de la méthode se fait de manière simple, en cliquant sur le nom correspondant. Une fois la liste de mots clés du corpus extraite, nous pouvons procéder au dépouillement en vue de l'identification des termes significatifs du point de vue terminologique. Pour le corpus français, le logiciel a automatiquement extrait une liste de 427 mots ; la liste des vingt premiers mots clés à l'état brut se présente de la manière suivante :

sécurité	programme
système	données
vulnérabilité	information
utilisateur	contrôle
accès	intégrité
risques	menaces
failles	peuvent
attaque	confidentielles
attaquant	privilèges
sensible	protéger

Nous constatons que contrairement aux listes de fréquences, la liste de mots clés n'inclut pratiquement pas de mots outils ; les tris seront donc plus rapides à faire. Même si la préparation du corpus de référence peut paraître une opération fastidieuse, ce travail préalable permet d'utiliser la fonction mots clés dans *AntConc* et donc de récupérer de manière très rapide une liste fiable de candidats termes. Il est bon de repréciser ici que la contribution du traducteur reste essentielle lorsqu'il s'agit de trier les unités retenues et de valider ou d'écarter les candidats termes extraits grâce à la fonction *Keyword List*. Dans la liste ci-dessus, le traducteur devra intervenir pour écarter au moins deux termes : « attaquant » et « peuvent ».

Nous pouvons décider d'extraire seulement 200 mots clés, par exemple. Pour cela, il faut définir le seuil de 200 termes et, dans notre cas, les 227 mots de la fin de la liste seront

écartés. Deux options d'analyse sont possibles : nous pouvons considérer que les mots qui se situent relativement bas dans la hiérarchie, après un seuil donné, sont moins significatifs pour notre recherche terminologique ; la deuxième possibilité consiste à parcourir toute la liste et à faire le tri manuel des candidats termes, en supposant que des termes intéressants puissent apparaître à la fin de la liste. Dans la liste de 427 mots clés extraits automatiquement dont il est question ici, les dix dernières entrées sont les suivantes :

prendre	composant
voire	contenant
sont	raison
augmentation	termes
chaque	vol

Ces termes candidats seront tous écartés, ce qui permet de penser que la détermination d'un seuil peut être une option intéressante pour limiter le travail de sélection que le traducteur devra faire sur les listes. Le seuil de 200 mots clés pour un corpus de 10 000 mots semble être un seuil assez élevé. Après le tri des mots candidats, le traducteur pourra retenir une liste de 150 à 160 mots pour le glossaire de spécialité. N'oublions pas que pendant le travail de tri, tous les mots clés de la liste sont actifs dans *AntConc*, ce qui veut dire qu'en cliquant sur un mot, nous accédons à sa concordance pour le voir en contexte et pouvons récupérer ainsi un certain nombre de syntagmes tels que : « accès par mandat », « vulnérabilité distante », « faille de sécurité », « attaque par messagerie », « attaque par débordement », pour ne mentionner que les candidats identifiés par le biais des concordances des premiers termes de la liste de mots clés.

Élaboration d'un glossaire bilingue

Le travail de sélection se fait sur la liste des mots clés extraits avec *AntConc*. Concrètement, le tri consiste à parcourir la liste et à éliminer les termes qui ne relèvent pas du domaine de la sécurité informatique. Pour juger de la pertinence d'un terme, le traducteur peut s'appuyer sur les concordances qui permettent de voir chaque terme en contexte ainsi que sur d'autres documents de spécialité auxquels il a accès. Le dépouillement des listes de mots clés et l'étude de leurs contextes nous a permis d'identifier 185 termes (mots et syntagmes, comme démontré ci-dessus) pour le corpus français. Nous avons procédé de la même manière pour le corpus anglais, pour lequel nous avons pu extraire 221 termes.

À partir des termes identifiés suite au travail d'extraction, un glossaire bilingue a été élaboré. Nous avons commencé le travail de construction du glossaire à partir de la liste des termes anglais. Pour faciliter le travail de mise en forme du glossaire, nous avons utilisé un tableur avec

nos listes de termes en parallèle, sur deux colonnes distinctes, et une colonne vide au milieu. Après avoir classé les termes de chaque liste par ordre alphabétique, nous avons procédé à des recherches dans la liste de termes français, en plaçant chaque terme identifié dans la colonne vide du milieu, à côte du terme anglais correspondant. À mesure que cette colonne se remplit, le glossaire prend forme. Nous avons aligné les termes de manière à pouvoir intégrer ce glossaire dans *Wordfast* et à l'utiliser comme base terminologique. Ce glossaire pourra bien évidemment être enrichi par la suite avec des termes récupérés dans de nouveaux corpus ou bien dans les textes que nous aurons à traduire. Il est mis en forme selon les spécifications de l'outil terminologique de *Wordfast* ; le glossaire au format .txt que nous constituons à partir des données du tableur se présente de la manière suivante :

access	accès
access control	contrôle d'accès
account	compte
account settings	paramètres du compte
antivirus	antivirus
attack	attaque
attempt	tentative, tenter
authentication	authentification
automated	automatisé
backdoor	porte dérobée
backup	sauvegarde

Analyse des résultats

En croisant les deux listes de mots clés, nous avons obtenu un taux de remplissage du glossaire d'environ 70 %. Cela revient à dire que pour 70 % des mots clés anglais retenus, la liste de mots clés français présentait les termes correspondants. Pour ce qui est des 30 % de termes restants, nous avons complété le glossaire en recherchant les équivalents français qui n'ont pas pu être récupérés dans la liste de mots clés. Nous avons fait appel à des dictionnaires en ligne, ainsi qu'à des recherches contextuelles dans des textes de spécialité. La même méthode a été appliquée aux 30 % de mots français restés sans équivalent après le croisement des deux listes de mots clés. L'ajout de termes permet d'affiner les glossaires ; cependant, le choix d'écarter les 30 % de termes qui sont restés sans équivalence peut également être envisagé. Les essais que nous avons réalisés montrent que le taux de correspondance de 70 % que nous avons obtenu est un résultat très satisfaisant et qu'il découle du degré de spécialisation des corpus analysés et de leur relative équivalence, bien qu'il ne s'agisse pas de corpus composés de traductions. Les différents essais pratiqués lors de la rédaction de cet

article montrent que l'extraction terminologique automatique à partir de corpus de traduction est plus rapide et plus fiable, avec des taux de correspondance entre les deux listes de termes (français et anglais) pouvant aller jusqu'à 95 %.

Conclusion

Comme nous avons pu le voir, l'extraction terminologique automatique à partir d'un corpus de textes appartenant à un domaine technique précis peut être réalisée de manière simple et accessible, grâce aux outils développés en TAL (traitement automatique des langues). Les fonctions d'un simple concordancier suffisent pour générer des listes de fréquences et des listes de mots clés. Bien que l'intervention du traducteur pour le tri des candidats termes et la constitution d'un glossaire bilingue représente une charge de travail non négligeable que le traducteur ne peut pas toujours se permettre lorsqu'il commence un nouveau projet de traduction, cette méthode porte ses fruits lorsque les glossaires sont utilisés et enrichis de manière régulière.

Bien évidemment, cette méthode d'extraction terminologique peut être appliquée à un texte que le traducteur aura à traduire, car elle permet de cibler les termes de spécialité, d'avoir un aperçu de leurs fréquences, de définir un vocabulaire spécifique et de préparer des glossaires qui seront réutilisables par la suite.

vasilica.le-floch@univ-lorraine.fr

Bibliographie

ANTHONY Laurence, 2011, *AntConc (Windows, Macintosh OS X, and Linux), fichier de documentation,* http://www.antlab.sci.waseda.ac.jp/software/README_AntConc3.2.4.pdf, consulté le 10 mars 2013.

ANTHONY Laurence, 2011, *AntConc*, http://www.antlab.sci.waseda.ac.jp/software.html, consulté le 10 mars 2013.

L'HOMME Marie-Claude, 2000, *Évaluation de logiciels d'extraction de terminologie : examen de quelques critères*, Montréal, Université de Montréal, http://olst.ling.umontreal.ca/pdf/terminotique/jiamcatt.pdf, consulté le 10 mars 2013.

LAROCHE Audrey, DROUIN Patrick, BERNIER-COLBORNE Gabriel, 2011, « Étude de l'influence de la taille du corpus de référence sur l'extraction terminologique automatique contrastive », Université de Montréal, http://olst.ling.umontreal.ca/pdf/Laroche_et_al_2011.pdf, consulté le 10 mars 2013.

LAVAULT-OLLEON Élisabeth, (éd.), 2007, *Traduction spécialisée : Pratiques, théories, formations*, Oxford, Peter Lang Publishing.

MEILLAND Jean-Claude, BELLOT Patrice, 2005, « Extraction automatique de terminologie à partir de libellés textuels courts », in WILLIAMS Geoffrey, (éd.), *Linguistique de corpus*, Rennes, Presses Universitaires de Rennes, pp. 357-370.

VAN CAMPENHOUDT Marc, LINO Teresa, COSTA Rute, (éd.), 2011, *Passeurs de mots, passeurs d'espoir*, Paris, Editions des Archives Contemporaines et Agence universitaire de la Francophonie.

Vasilica Le Floch *est traductrice assermentée auprès de la Cour d'Appel de Metz et ses langues de travail sont l'anglais, le français et le roumain. Ses domaines de spécialité sont la traduction juridique et la traduction technique, notamment dans le secteur de l'informatique. Titulaire d'un Doctorat de Linguistique anglaise, elle enseigne l'anglais de spécialité à l'Université de Lorraine, IUT Nancy Charlemagne. Elle est membre de l'équipe de recherche IDEA, Interdisciplinarité dans les Études Anglophones. La traduction constitue l'un de ses domaines de recherche.*