TERMINUS
Station de travail de gestion de la terminologie et des corpus

1. Introduction

Dans cet article, nous présenterons un logiciel spécialement conçu pour la gestion des corpus et de la terminologie, selon les grandes lignes de la théorie communicative de la terminologie (Cabré, 1999), abrégée ci-après TCT. La TCT propose une approche linguistique de la terminologie, prenant également en compte des aspects cognitifs et discursifs ; son objet d'étude porte donc sur des termes pris dans le contexte d'un discours spécialisé. Dans le cadre de la TCT, l'approche linguistique de la terminologie se fait à partir de corpus, rejoignant les partisans de la linguistique de corpus (Sinclair, 1991 ; Sager, 1991 ; McEnery et Wilson, 1996 ; etc.). Cela signifie que la TCT ne s'intéresse pas uniquement à la terminologie prescriptive, c'est-à-dire celle qui est établie par des normes ou issue de bases de données officielles, mais aussi (et particulièrement) aux termes des corpus de langue spécialisée (LSP) utilisés par des experts. Autrement dit, la TCT implique non seulement une approche *in vitro*, mais s'intéresse également aux termes *in vivo*.

La mise au point du logiciel TERMINUS est le fruit des efforts conjugués de plusieurs linguistes et programmeurs de l'Institut Universitari de Lingüística Aplicada de l'Université Pompeu Fabra, à Barcelone. TERMINUS est une plateforme Web qui fournit aux utilisateurs des outils permettant d'effectuer l'ensemble des démarches d'un travail de terminologie, telles que la constitution et l'analyse de corpus, l'extraction de termes ou encore la création d'une base de données pour stocker et récupérer des fiches terminologiques.

Les raisons qui nous ont conduits à créer ce programme sont liées à notre expérience de l'enseignement de la terminologie, notamment en ce qui concerne l'élaboration de projets terminologiques multilingues à partir de corpus. En effet, nous nous sommes aperçus qu'il y avait un décalage entre les idées que nous transmettions à nos étudiants et ce qu'ils pouvaient faire, dans la pratique, avec les outils actuellement disponibles sur le marché. De même, nous avons également remarqué que l'exécution de ces projets implique des notions trop techniques ou abstraites pour qu'on puisse les expliquer uniquement par des mots. Par conséquent, nous avons décidé de mettre au point un outil informatique qui permette aux étudiants de mener leurs propres expériences afin de mieux saisir certaines notions complexes, ou, autrement dit, d'*apprendre en faisant*.

2. Fondements de la terminologie basée sur des corpus

Le but de la théorie de la terminologie est de produire une description formelle, sémantique et fonctionnelle des unités lexicales ayant une valeur terminologique, et d'expliquer leurs rapports avec les autres unités du système linguistique. La TCT, comme il a été dit précédemment, s'intéresse aux termes *vivants*, à savoir les unités qui sont effectivement utilisées par des experts dans des communications spécialisées. En tant que théorie linguistique des termes, la TCT conçoit son objet (c'est-à-dire les termes) comme des unités cognitives exemplifiées en tant qu'unités lexicales dans le langage naturel. Il s'agit d'une théorie multidimensionnelle et variationniste qui met l'accent sur l'aspect communicatif de l'utilisation des termes, au point qu'elle propose de passer d'une théorie basée sur la dénomination à une théorie de la communication. L'unité terminologique est donc conçue comme un polyèdre à trois faces, formé d'un composant cognitif (le concept), d'un composant linguistique (le terme) et d'un composant communicatif (la situation). Ainsi, la terminologie n'apparaît plus comme un champ autonome, mais interdisciplinaire, dans la mesure où les termes peuvent être l'objet d'étude d'une théorie du langage, d'une théorie de la connaissance et d'une théorie de la communication.

Par définition, toute approche de la terminologie basée sur des corpus vise à analyser des unités qui apparaissent de manière naturelle dans des textes spécialisés qui véhiculent des connaissances pointues sur un sujet donné. Comme nous l'avons déjà indiqué, notre travail porte sur la terminologie vivante et, en menant des analyses empiriques, l'on s'aperçoit que la terminologie est beaucoup plus complexe qu'on n'avait coutume de le croire. Tout comme la langue générale, les textes spécialisés contiennent aussi des cas de synonymie, des ambiguïtés, des flous, des périphrases, des redondances et des systèmes de variation de terme qui se produisent à différents niveaux. Devant ce constat, la notion d'univocité (selon laquelle les concepts sont anhistoriques et conçus comme des entités indépendantes du langage, dont chacune correspondrait à un terme particulier dans différentes langues) devient indéfendable. Par conséquent, toute théorie de la terminologie doit partir du principe que la variation est une propriété essentielle de la communication entre experts, et que termes et concepts doivent être étudiés dans leur interaction dynamique.

3. Analyse empirique

Comme il a été annoncé dans l'introduction, l'analyse empirique des corpus sera réalisée au moyen du logiciel TERMINUS, une application Web conçue pour l'analyse de corpus et pour l'extraction et la gestion de termes (voir capture d'écran de la page d'accueil du programme en figure 1). À l'aide de ce programme, l'on pourra effectuer toutes les démarches d'un projet terminologique type, dont les plus importantes sont expliquées dans ce chapitre, en vue d'aboutir à la création d'un glossaire, qui sera le produit final de ce projet.

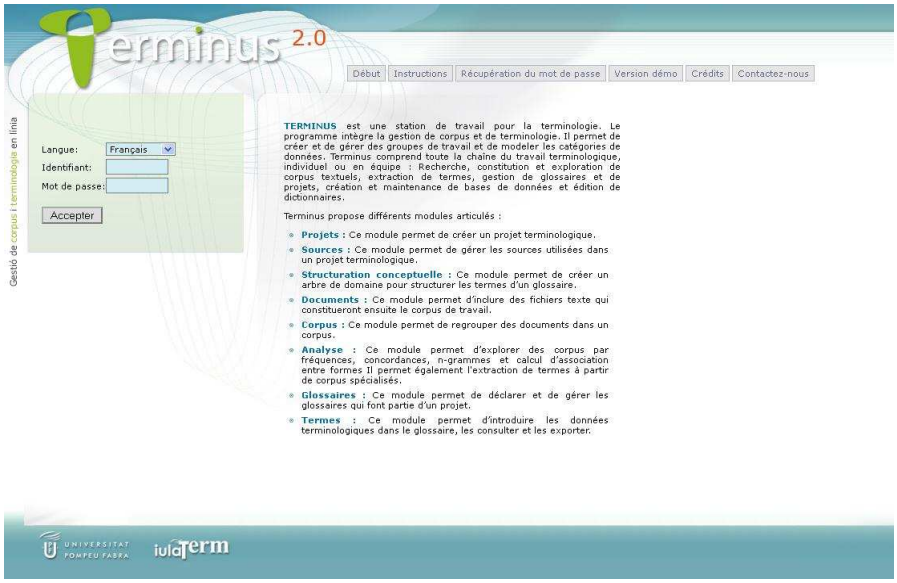

Figure 1. Page d'accueil de TERMINUS

La première étape consiste à définir le projet terminologique. Lors de cette phase, un certain nombre de choix importants doivent être faits : le domaine des textes à analyser en vue de créer un glossaire ; la langue principale utilisée pour les entrées ; la langue ou les langues des termes équivalents ; et surtout, le public auquel s'adresse le produit terminologique (situation communicative). Une fois ces décisions prises, l'étape suivante consiste à élaborer un corpus sur lequel on puisse travailler, c'est-à-dire un échantillon de documents spécialisés suffisamment consistants et de bonne qualité pour être considérés comme représentatifs du domaine choisi.

3.1. Constitution d'un corpus

La constitution d'un corpus de documents poursuit un triple objectif. Premièrement, en tant que terminologues, nous avons besoin de nous frotter directement au type de langage utilisé dans le domaine en question pour nous familiariser avec celui-ci ; une expérience indispensable qui complètera les informations que l'on aura obtenues en interrogeant les experts. Deuxièmement, ce corpus est nécessaire pour pouvoir mener différentes analyses statistiques du vocabulaire et procéder à l'extraction de termes. Troisièmement, ces textes permettront d'obtenir des informations complémentaires sur les termes utilisés, notamment au niveau de la sémantique, de la syntaxe et de la collocation.

Dans l'idéal, le corpus à analyser devrait être suffisamment consistant pour pouvoir être considéré comme représentatif du domaine. Malheureusement, il n'existe aucune formule mathématique précise qui permette de déterminer le volume que doit avoir un corpus pour que l'on puisse garantir sa représentativité (il existe cependant quelques méthodes approximatives ; voir par exemple Biber, 1993). En général, les spécialistes de la linguistique de corpus affirment que *« plus il y a de matière, mieux c'est »*. Un corpus doit donc réunir le plus de documents possibles, car plus il sera volumineux, plus il contiendra d'unités terminologiques, et plus fiables seront nos conclusions.

Hormis le volume du corpus, la question la plus importante concerne sans doute l'aspect qualitatif. Quel doit être le corpus idéal du domaine que l'on étudie ? Lorsqu'on aborde un nouveau domaine de connaissances, la première démarche consiste tout d'abord à repérer les publications de référence. Où les experts publient-ils leurs documents ? Il s'agit sans doute de la meilleure façon de constituer un corpus de langue spécialisée ; mais peut-être pas la plus pratique. En effet, il arrive parfois que les documents ne soient pas disponibles au format électronique ; or les coûts d'une numérisation par ROC et des corrections manuelles à effectuer par la suite représentent une somme colossale. Dans ce cas, il s'avère peut-être préférable de lancer une recherche de documents sur la Toile en entrant certains termes en guise d'expression de requête. Pour constituer un corpus, avec TERMINUS, l'utilisateur a le choix entre deux formules : soit il charge sur la plateforme les fichiers qu'il a recueillis au préalable ; soit il sélectionne l'une des options de TERMINUS qui permettent de télécharger des documents à l'aide de moteurs de recherche sur Internet. Cette seconde option peut être utile si l'on souhaite télécharger une quantité importante de données, mais il est fortement recommandé de les examiner manuellement par la suite, afin de supprimer les documents non pertinents.

3.2. Création d'une structure conceptuelle

Lors des premières phases de l'analyse, il s'avérera peut-être utile de développer la structure conceptuelle du domaine étudié, dans la mesure où cet outil peut aider l'analyste, qui en général n'est pas un expert de la matière traitée, à établir une hiérarchie claire des concepts inhérents à ce domaine. En d'autres termes, la structure conceptuelle illustre le processus par lequel le/la terminologue acquiert des connaissances sur le domaine qu'il/elle étudie.

Parmi ses principales fonctions, TERMINUS possède un module de structuration conceptuelle qui permet de dessiner l'arbre conceptuel du domaine étudié (figure 2). Cet arbre est composé de boîtes qui contiennent les termes utilisés dans ce domaine. Les liens entre les boîtes reflètent les rapports qui existent entre les différents concepts. Ces informations, introduites par l'utilisateur sous forme de représentation graphique, seront ensuite encodées par le programme sous forme de représentation logique (en syntaxe XML), de façon à ce qu'elles puissent être traitées par d'autres systèmes.

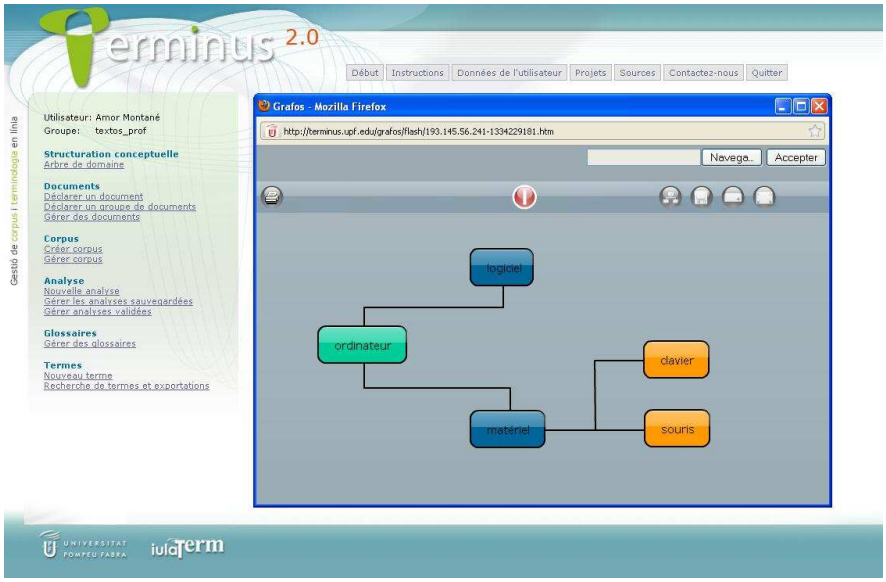

Figure 2. Module de structuration conceptuelle de TERMINUS

3.3. Analyse d'un corpus

Une fois que l'utilisateur a défini et constitué son corpus, TERMINUS propose différentes méthodes pour analyser le vocabulaire. Chacune sera décrite en détail dans les sous-chapitres qui suivent : extraction de concordances (ou mots-clés en contexte), expliquée au point 3.3.1 ; classification du vocabulaire (mots ou n-grammes) par fréquence ou par mesures d'association statistiques, point 3.3.2 ; et extraction automatique de terminologie, point 3.3.3.

3.3.1. Extraction de concordances

La recherche de mots-clés en contexte, aussi appelée *extraction de concordances* dans le champ de la linguistique de corpus, est une des méthodes d'analyse de corpus les plus simples qui existent. Celle-ci consiste à extraire les contextes d'occurrence d'un terme ou d'une expression donnée, ces contextes pouvant être une phrase ou un nombre de mots arbitraire situés à gauche et à droite du terme en question. L'extraction de concordances peut être extrêmement utile pour cerner rapidement la signification d'un terme, en observant la façon dont les experts l'utilisent dans des textes réels. On peut également avoir recours à cette méthode pour analyser les collocations du terme ou pour voir les autres termes auxquels il est associé sur le plan conceptuel.

Figure 3. Extraction de concordances

3.3.2. Classification du vocabulaire du corpus

La classification du vocabulaire est une autre stratégie qui permet d'analyser un corpus. À cet effet, le programme offre la possibilité de classer des mots ou des séquences de mots (n-grammes) par ordre alphabétique, par ordre de fréquence décroissante ou en utilisant des mesures d'association statistiques qui signalent les séquences de mots apparaissant ensemble, et ce de façon significative, au sein du corpus.

Ce procédé extrêmement simple permet de repérer des termes polylexicaux, des collocations et des unités phraséologiques de différentes sortes. Pour l'extraction automatique de terminologie, cependant, TERMINUS est doté d'une fonction spécifique, qui sera décrite au point suivant.

3.3.3. Extraction de termes

L'extraction automatique de terminologie est un des domaines de recherche les plus dynamiques de la terminologie. Ce procédé consiste à élaborer des algorithmes informatiques afin d'extraire des unités terminologiques à partir des textes rassemblés. Il s'agit d'un domaine d'expertise à part entière, extrêmement technique, et malgré les efforts réalisés au cours des dernières décennies (voir par exemple Justeson et Katz, 1995 ; Kageura et Umino, 1998 ; Bourigault et al., 2001), à notre connaissance, il n'existe toujours pas de consensus sur la meilleure stratégie à adopter pour l'extraction de termes. Cette question reste donc en suspens.

Pour procéder à cette opération, TERMINUS est doté d'un algorithme original qui lui est propre. À l'instar de la plupart des extracteurs de termes, les résultats doivent ensuite être validés par l'utilisateur, car les candidats-termes extraits par le système ne sont pas toujours de véritables termes. Toutefois, il ne fait aucun doute que les résultats de cet extracteur de termes sont d'une qualité supérieure à ce que l'on pourrait obtenir en utilisant les méthodes d'extraction décrites au point 3.3.2.

Figure 4. Extraction de termes

Nous n'examinerons pas en détail le fonctionnement du logiciel, car ce n'est pas notre propos dans cet article (pour en avoir une description, voir Cabré et Nazar, à paraître). Globalement, cependant, l'on peut dire que c'est un programme qui « apprend » à partir des exemples fournis par les utilisateurs. Lors d'une phase préalable à l'analyse, l'utilisateur doit « instruire » le programme en y incorporant des listes de termes relatifs au domaine étudié, dans une langue donnée. À partir de cette liste, le programme développera un modèle mathématique concernant ces termes. Une fois le programme instruit, celui-ci sera prêt à extraire un nombre indéfini de termes liés au domaine en question. L'utilisateur peut affiner le processus d'apprentissage du logiciel en l'enrichissant peu à peu de nouveaux termes extraits du corpus. Étant donné qu'il s'agit d'une application Web, tout ce que TERMINUS apprend d'un utilisateur est mis en partage, au bénéfice de tous les utilisateurs, ce qui est l'atout le plus intéressant de ce programme. Ainsi, le programme reste en apprentissage constant, grâce aux contributions de la communauté de terminologues, et peut s'adapter à d'autres langues et domaines sans qu'il ne faille modifier le logiciel en lui-même.

3.4. Création d'un glossaire

Créer un glossaire implique toute une série d'importantes décisions. TERMINUS possède un modèle de glossaire intégré qui comporte les champs que l'on trouve le plus fréquemment dans un glossaire, notamment la catégorie grammaticale du terme, sa source, les contextes d'occurrence, les équivalences, les collocations, et bien d'autres informations. De plus, l'utilisateur peut configurer son propre glossaire en personnalisant, en supprimant ou en créant de nouveaux champs. Cette fonctionnalité vise à fournir un minimum d'instructions aux utilisateurs novices qui souhaitent créer un glossaire, tout en s'adaptant au maximum aux exigences des utilisateurs experts,

de façon à ce que leurs glossaires puissent répondre à leurs besoins spécifiques ou à ceux des utilisateurs finals ciblés.

3.5. Gestion des termes

Lorsque les champs du glossaire ont été définis et que le programme a été configuré en conséquence, il est possible de créer une fiche terminologique où l'on fera apparaître des informations sur chaque terme. Une fois les fiches complétées, il ne reste plus qu'à exporter le glossaire dans un fichier selon l'un des différents formats disponibles. Pour des lecteurs humains, les formats PDF et HTML sont les plus pratiques (la figure 5 montre un exemple de fiche terminologique en HTML). En revanche, les utilisateurs qui souhaitent exporter les données, en vue de les importer par la suite dans un autre logiciel de base de données, préféreront peut-être d'autres formats, tels que XML ou CSV.

Figure 5. Résultat de la recherche de termes

4. Conclusions

L'objet de cet article, destiné à la communauté de professionnels en rapport avec la terminologie, était de présenter un outil récemment mis au point, conçu pour les travaux de terminologie à partir de corpus. Terminus est un système intégral permettant de réaliser l'ensemble des démarches inhérentes à de tels travaux : de la constitution d'un corpus à l'édition d'un glossaire, en passant par l'analyse et l'exploitation de données textuelles et terminologiques, ainsi que l'élaboration d'une structure conceptuelle aidant à sélectionner et à organiser les termes répertoriés dans le glossaire. Toutes ces opérations ont été brièvement décrites dans cet article.

Pour conclure, nous aimerions souligner le fait que TERMINUS est un système modulaire. Celui-ci est en effet constitué de modules indépendants qui peuvent être utilisés séparément. Cet outil conviendra donc à des utilisateurs ayant des profils et

des besoins fort différents les uns des autres. Terminus s'avère d'autant plus adaptable que les champs des fiches terminologiques peuvent être personnalisés de façon à répondre aux besoins les plus variés en matière de terminologie.

Références

BIBER, D. « Representativeness in Corpus Design ». In *Literary and Linguistic Computing,* vol. 8, n° 4, 1993, p. 243-257

BOURIGAULT, D. ; JACQUEMIN, C. ; L'HOMME, M-C. (éd.). *Recent Advances in Computational Terminology*. Amsterdam : John Benjamins, 2001

CABRÉ, M. T. *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelone : Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, 1999

CABRÉ, M. T. « Theories of terminology. Their description, prescription and explanation ». In *Terminology*, vol. 9, n° 2. Amsterdam : John Benjamins, 2003, p. 163-200

CABRÉ, M. T. ; NAZAR, R. « Supervised Learning Algorithms Applied to Terminology Extraction ». In *Proceedings of TKE 2012*, Madrid, 20-21 juin 2012 (à paraître)

JUSTESON, J. ; KATZ, S. « Technical terminology: some linguistic properties and an algorithm for identification in text ». In *Natural Language Engineering*, vol. 1, n° 1, 1995, p. 927

KAGEURA, K. ; UMINO, B. « Methods of Automatic Term Recognition ». In *Terminology*, vol. 3, n° 2. Amsterdam : John Benjamins, 1998, p. 259-289

MCENERY, T. ; WILSON, A. *Corpus Linguistics: An Introduction*. Édimbourg : Edinburgh University Press, 1996

SINCLAIR, J. *Corpus, concordance, collocation*. Oxford : Oxford University Press, 1991