Fusion d'un thesaurus et d'une terminologie : utilisation de ressources existantes pour amorcer une onto-terminologie

Laurence Kister[1], Evelyne Jacquey[2], Bertrand Gaiffe[2]

[1] UMR ATILF – Nancy Université – CNRS,
Laurence.Kister@univ-nancy2.fr
[2] UMR ATILF – CNRS – Nancy Université,
{Evelyne.Jacquey, Bertrand.Gaiffe}@atilf.fr

Résumé : Cet article porte sur l'enrichissement d'un thesaurus par une terminologie extraite d'un dictionnaire. Le thesaurus est structuré tandis que la terminologie apporte des définitions. La fusion des deux ressources par enrichissement mutuel permet de s'orienter vers une onto-terminologie destinée à l'annotation sémantique de textes de spécialité. Nous examinons ici les difficultés liées à structuration d'une terminologie en onto-terminologie et à celles de l'enrichissement d'un thésaurus.

Mots-clés : Terminologie, Acquisition de ressources, Etiquetage de textes, Classification, Enrichissement de textes

Abstract : In this paper, we are interested here in paper enrichment of a thesaurus using a terminology extracted from a dictionary. The thesaurus is structured and the terminology brings definitions. The merger of both resources in an onto-terminology is intended to contribute to the semantic tagging of texts of speciality. We examine the difficulties related to structuring of a terminology in an onto-terminology and those related to the enrichment of a thesaurus.

Keywords : Terminology, Resources acquisition, Texts tagging, Classification, Texts enrichment

1 Introduction

L'augmentation et la diversification des échanges conduisent à une explosion de la quantité d'informations dans les différents domaines scientifiques et à une interpénétration des domaines en fonction d'approches multi- ou interdisciplinaires. Les ressources terminologiques et ontologiques (Bourigault & Aussenac-Gilles, 2003) sont utilisés dans différents domaines : information et documentation (Roche, 2004), édition (El Mekki & Nazarenko, 2002), recherche d'informations, la classification de documents (Sanjuan & Ibekwe-Sanjuan, 2002). L'extraction automatique et la structuration de ressources terminologiques ont donné lieu à la réalisation d'outils fondés sur des méthodes symboliques et numériques, utilisant

différents types d'informations pour leur structuration (relations hiérarchiques d'ordre conceptuel, relations de sémantique lexicale, analyse distributionnelle des contextes d'apparition des termes dans les textes, etc.) ((Nazarenko & Hamon, 2002), (Bourigault et *al.*, 2001) et (L'Homme, 2004)). Dans ces travaux, les ressources terminologiques sont extraites de textes de spécialité (Dell'Ortella & *al.*, 2008). L'utilisation de ressources externes, tels des dictionnaires, est seconde.

Nous partons d'une terminologie des sciences du langage extraite du TLFi fortement domanialisé[1], validée sur un corpus spécialisé en sciences du langage (Gaiffe & *al.*, 2009). La terminologie extraite est suffisamment riche sémantiquement pour différencier les emplois terminologiques des emplois de langue courante. En revanche, sa structure est plate. Par ailleurs, nous disposons du thesaurus Thesaulangue[2], mis au format TMF par l'Inist dans le cadre du projet Termsciences. Bien que structuré, il offre une image incomplète de la linguistique actuelle : sa structure reste marquée par l'époque de son initialisation (Inalf, 1975). Les développements récents et les usages non stabilisés de l'époque font que : (1) un terme peut renvoyer à plusieurs termes dans différents sous-domaines : *morphologie lexicale* renvoie à *lexicologie* et *sémantique* sans plus de précisions alors qu'il s'agit d'intitulés d'étiquettes de micro-thesaurus, (2) des termes renvoient à d'autres termes qui ne font l'objet d'aucun développement : *analyse du discours* est employé pour *analyse textuelle, linguistique textuelle* et *praxématique* et renvoie à *informatisation de corpus textuels* – un intitulé du micro-thesaurus non développé, (3) certains termes sont employés pour plusieurs autres et renvoient à des termes relevant de plusieurs micro-thesaurus : *industrie de la langue* est utilisé pour *ingénierie linguistique* et *génie linguistique* et renvoie vers *banque de données, informatisation de corpus textuels, logiciel, informatisation de dictionnaires, traduction automatique* qui appartiennent aux micro-thesaurus *approches interdisciplinaires* pour les trois premiers, *lexicologie* pour le quatrième et *linguistique appliquée* pour le dernier, (4) certains termes font l'objet de renvois réciproques dans un même micro-thesaurus : *conjonction de coordination* et *coordination* apparaissent tous deux en *syntaxe* dans des sous-domaines distincts (*syntaxe => conjonction => conjonction de coordination => adversative* vs *syntaxe => phrase => construction => coordination*) et renvoient l'un à l'autre.

Malgré les imperfections Thesaulangue constitue un bon point de départ pour la structuration de la terminologie plane extraite du TLFi. En effet, les documentalistes qui sont intervenus et qui font encore régulièrement évoluer le thésaurus ont pallié à certaines anomalies en introduisant et développant de nouvelles subdivisions, parfois en parallèle (*cf.* le cas de *syntaxe* en 4.1). Les micro-thesaurus les plus développés constituent une base solide pour la construction d'une onto-terminologie destinée à faciliter l'indexation du contenu des documents, à révéler les usages des termes du domaine de spécialité dans les textes, à faire apparaître les évolutions sémantiques

[1] Le Trésor de la Langue Française informatisé (TLFi) [Dendien & Pierrel, 2002] - comporte de nombreux sens explicitement associés à des domaines de spécialité : 97330 sens domanialisés sur 271165, soit près de 36 %.

[2] Thesaulangue est le thésaurus mis au point et maintenu par le centre de documentation de l'atilf – débuté par l'Inalf - qui a été intégré au portail Termsciences de l'Inist.

ainsi que la répartition des termes en fonction du niveaux de scientificité ou de vulgarisation du document (Kister & Jacquey, 2007a et b - Kister & *al.*, 2008). Notre approche rejoint celle des auteurs qui ont montré la possibilité d'indexer des textes à partir d'une terminologie ou d'un thésaurus (Bourigault & *al.*, 2004) et (Aussenac-Gilles & Bourigault, 2000).

2 Les données en présence

La terminologie des *sciences du langage* a été extraite à partir du TLFi en utilisant 12 des sous-domaines des sciences du langage utilisés pour classer les entrées en grandes disciplines scientifiques et techniques[3] : *grammaire, lexicographie, lexicologie, linguistique, philologie, phonétique, phonologie, rhétorique, sémiologie, sémiotique, stylistique, toponymie*. Elle comporte 2042 candidats termes caractérisés par des informations lexicographiques : conditions d'emploi, liens de synonymie éventuels, syntagmes illustratifs du sens domanialisé et un ou plusieurs exemples.

Le thesaurus Thesaulangue comporte 844 termes structurés selon 20 micro-thesaurus rattachés à la racine *sciences du langage* auxquels s'ajoutent 335 termes pour lesquels on conseille l'emploi des 844 termes structurés, soit 1179 termes.

L'intersection des deux ressources - pour cette première expérience nous avons confrontés les 844 termes utilisés pour l'indexation - concerne 307 termes.

3 Résultats de la fusion des deux ressources

La fusion conduit à un résultat satisfaisant qui permet d'enrichir une partie des termes du thesaurus avec des définitions, mais plusieurs types d'interrogation découlent de cette expérience : (1) la nécessité de compléter le thesaurus par adjonction des termes définis dans la terminologie qui ne sont pas dans le thesaurus, (2) la possibilité de définir les termes du thesaurus qui ne figurent pas dans la terminologie extraite du TLFi, (3) l'importance à accorder à la structure du thesaurus, (4) la prise en compte ou non des informations domaniales utilisées pour extraire la terminologie du TLFi.

3.1 Divergences entre les termes dominants de Thesaulangue et du TLFi

Différents types de divergences surviennent entre le terme dominant de Thesaulangue ('broaderGenericConcept' dans le format TMF) et le domaine extrait du TLFi : (1) 188 termes ont un domaine générique dans le TLFi, (2) 82 termes de la *syntaxe* dans Thesaulangue sont domanialisés, par *grammaire* dans le TLFi, (3) 29 termes de Thesaulangue sont caractérisés par un domaine du TLFi non conforme à la

[3] Les domaines et les sous-domaines utilisés dans le TLFi sont accessibles à partir de la recherche assistée. Les domaines des sciences du langage non retenu car jugés trop marginaux ou correspondant à un domaine indépendant sont : *épigraphie, graphologie, paléographie, terminologie, versification*.

structure du thesaurus, (4) 3 termes de Thesaulangue sont porteurs d'un domaine TLFi différent mais compatible.

3.2 Etiquetage domanial dans le TLFi et dans Thesaulangue

3.2.1 Le domaine du TLFi est trop général

Près de 10% des cas posent la difficulté d'un positionnement dans un domaine différent dans le TLFi et dans Thesaulangue. Les domaines utilisés par le TLFi restent généraux : ils sont utiles pour extraire les termes du domaine des *sciences du langage* et les informations lexicologiques qui s'y rapportent mais ils manquent de précisions pour donner une représentation structurée du domaine. Si on prend l'exemple des termes *syntagmatique, pléonasme, dérivation impropre*, ils sont respectivement classés dans les domaines *sémiotique*, *linguistique stylistique* et *lexicologie* du TLFi alors qu'ils sont positionnés au troisième et au quatrième niveaux hiérarchiques dans Thesaulangue : (1) *lexicologie => rapports => syntagmatique, sémantique => rhétorique => figure de style => **pléonasme***, (2) *morphologie => morphologie lexicale => dérivation => **dérivation impropre***.

Certains termes relevant de la *sémiotique* dans Thesaulangue sont répartis entre la *grammaire* et la *linguistique* dans le TLFi. Pour la *syntaxe* certains termes de Thesaulangue figurent sous *sémiotique, linguistique structurale, grammaire générative* et *sémantique contemporaine* du TLFi. Deux de ces quatre domaines figurent dans Thesaulangue : *sémiotique* est un micro-thesaurus et *grammaire générative* un sous-domaine (*méthodologie linguistique => fonctionnalisme => grammaire générative*). Les deux autres domaines - peu utilisés dans le TLFi - n'apparaissent pas dans Thesaulangue. Des termes relevant de la *sémantique* dans Thesaulangue sont étiquetés comme appartenant aux domaines *grammaire* et *sémiotique* du TLFi. Des termes du micro-thesaurus *morphologie* appartiennent au domaine *phonétique, lexicologie* et *grammaire* du TLFi. Pour l'ensemble des divergences l'utilisation de la structure du thesaurus permet d'atteindre un degré de précision supérieur : les approches et les disciplines des *sciences du langage* y sont plus détaillées et plus structurées. Ceci conforte notre démarche : l'utilisation de Thesaulangue comme fondement d'une onto-terminologie est possible même si des remaniements, le développement er la restructuration de certains micro-thesaurus s'imposent.

3.2.2 Termes trop généraux et double étiquetage dans le TLFi

La généricité des étiquettes de domaines est trop importante ce qui a pour conséquences une utilisation trop fréquente de certaines d'entre elles. Ces étiquettes sont aussi régulièrement utilisées simultanément à d'autres étiquettes relativement génériques qui ne permettent pas la structuration des concepts du thesaurus. Par exemple, pour *perfectif* nous ne retenons ni *linguistique* ni *grammaire* qui apparaissent pour l'entrée terminologique extraite à partir du TLFi, mais la hiérarchie de Thesaulangue (*syntaxe => aspect => verbe => perfectif*) que nous enrichissons

avec l'information lexicographique du TLFi (*perfectif* DEF= *(Aspect grammatical) qui envisage le procès dans son terme*).

3.3 Un exemple de désaccord : *grammaire* ou *syntaxe*

Dans le TLFi, les *sciences du langage* ne comportent pas le domaine *syntaxe* ce qui a eu pour conséquence de domanialiser ce qui relève de la syntaxe sous *grammaire*. Or, *syntaxe* correspond à une partie développée de Thesaulangue et par conséquent de Termsciences. En effet, le thesaurus comporte un micro-thesaurus *grammaire* (1 niveau de descendance composé de 9 fils : grammaire historique – études sur les grammaires – grammaires du Moyen âge ainsi que grammaires du 16ème, 17ème, 18ème, 19ème, 20ème et 21ème siècle) et un micro-thesaurus *syntaxe* (5 niveaux de descendance avec 20 fils pour le 1er niveaux, 58 pour le 2ème, 65 pour le 3ème, 37 pour le 4ème et 2 pour le 5ème soit un total de 180 étiquettes. Le choix de tout étiqueter *grammaire* dans le TLFi rend impossible la structuration et la hiérarchisation de l'information. Ainsi, même si on accepte de considérer la syntaxe comme un sous-domaine de la grammaire, on ne peut se résoudre à tout classer sous l'étiquette générique de grammaire qui ne correspond pas au générique de syntaxe dans le thesaurus : *syntaxe* est un fils de *sciences du langage* au même titre que *grammaire*. Pour ce genre de divergence, nous conserverons la structure de Thesaulangue. On peut alors s'interroger quand à la pertinence du micro-thesaurus *grammaire* ou de son articulation avec celui de la *syntaxe.*

4 Perspectives

L'état d'avancement de cette expérience laisse entrevoir l'ampleur de la tâche qu'il reste à effectuer pour produire une onto-terminologie dont les informations sémantiques pourront être reportées sur un document en texte intégral : (1) développer certaines parties du thesaurus par l'insertion des termes qui apparaissent dans la terminologie qui ne sont pas dans le thesaurus, (2) structurer la ressource qui résulte de la fusion des deux ressources préexistantes, (3) confronter les 335 termes non utilisés pour indexer les documents à la terminologie pour y associer d'éventuelles définitions afin d'offrir une représentation plus précise du domaine, (4) définir les termes du thesaurus non définis par la terminologie, (5) enrichir la ressource produite en traits sémantiques pour permettre l'annotation sémantique automatique ou semi-assistée, (6) valider la ressource produite pour vérifier que l'image qu'elle donne des *sciences du langage* est satisfaisante.

A plus long terme l'onto-terminologie peut constituer un outil pédagogique facilitant la connaissance et l'utilisation des termes et du thesaurus des *sciences du langage* par des documentalistes novices du domaine ou tout utilisateur le découvrant.

5 Références

AUSSENAC-GILLES N. & BOURIGAULT D. (2000). The Th[IC]2 Initiative : Corpus-Based Thesaurus Construction for Indexing WWW Documents. *Proceedings of the EKAW'2000 workshop « Ontologies and texts »*, Juan-Les-Pins, Université Paul Sabatier, Toulouse, p. 71-78, octobre.

BOURIGAULT D., JACQUEMIN C. & L'HOMME M.C. (2001). Recent Advances in Computational Terminology. Amsterdam/Philadelphie : John Benjamins.

BOURIGAULT D. & AUSSENAC-GILLES N. (2003). Construction d'ontologies à partir de textes. *Actes de la conférence TALN 2003*, Batz-sur-Mer.

BOURIGAULT D., AUSSENAC-GILLES N. & CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. In M. SLODZIAN (Ed.), *Revue d'Intelligence Artificielle (RIA), Numéro spécial sur les techniques informatiques de structuration de terminologies*, Hermès, Paris, Vol. 18, N. 1/2004, p. 87-110, 2004.

DELL'ORTELLA F., LENCI A., MARCHI S., MONTEMAGNI S., PIRELLI V. & VENTURI G. (2008). Dal testo alla conoscenza e ritorno : estrazione terminologica e annotazione semantica di basi documentali di dominio. *Analisi Testuale e Documentazione nella città digitale, Convegno nazionale dell'Associazione Italiana per la Terminologia*. I-TerAnDo, Università di Calabria, Rende, 5-7 juin, AIDAinformazioni, 26, 1-2, pp. 185-206.

DENDIEN J. & PIERREL J.M. (2002). Le trésor de la langue informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *TAL*.

EL MEKKI T. & NAZARENKO A. (2002). Comment aider un auteur à construire l'index d'ouvrage. *Actes de la conférence CIFT 2002*. pp. 141 – 157. Tunisie.

GAIFFE B., JACQUEY E. & KISTER L. (2009). Approche lexico-sémantique de l'extraction terminologique : utilisation de ressources lexicographiques et validation sur corpus. *Toth'09*, Annecy, 4-5 juin.

INALF. (1975). Trésor de la langue française. *Le français moderne*, (supplément), nouvelle série, fascicule 2.

KISTER L., JACQUEY E. & GAIFFE B. (2008). Repérage de la référence à partir du thesaurus, de la terminologie et de la sémantique lexicale. *Analisi Testuale e Documentazione nella città digitale*. Convegno nazionale dell'Associazione Italiana per la Terminologia, I-TerAnDo, Università di Calabria, Rende, 5-7 juin, AIDAinformazioni, 26, 1-2, pp. 25-36.

KISTER L. & JACQUEY E. (2007a). Comparaison des structures thématiques de textes spécialisés et de thésaurus ou de terminologies. *Terminologia e mediazione linguistica : approcci e metodi a confronto*. ASS.I.term et Università di Bologna, sede di Forli, Bertinoro, 8 juin, Realiter, en ligne, (http://realiter.net/spip.php?article951).

KISTER L. & JACQUEY E. (2007b). Acquisition sémantique à partir de données lexicographiques au service de la comparaison entre des structures thématiques de textes spécialisés et de thésaurus. *Terminologie : approches transdisciplinaires*, Gatineau (Québec), 2-4 mai, an ligne, (http://www.uqo.ca/terminologie2007/documents/kister_Jacquey.pdf).

L'HOMME M.C. (2004). La terminologie : Principes et Techniques. Presses Universitaires de Montréal.

NAZARENKO A. & HAMON T. (2002). Structuration de terminologie. *TAL*, vol 43, n°1, 174 pages.

ROCHE M. (2004). Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes. PhD Thesis, Université Paris 11.

SANJUAN E. & IBEKWE-SANJUAN F. (2002). Terminologie et classification automatique des textes, *Actes de la conférence JADT 2002*, pp. 677-688.