TOTh 09

Terminologie & Ontologie : Théories et Applications

Actes de la troisième conférence TOTh - Annecy - 4 & 5 juin 2009

Institut Porphyre
Savoir et Connaissance

Dans un monde où la communication et le partage d'information sont au cœur de nos activités, les besoins en terminologie se font de plus en plus pressants. Il est devenu impératif d'identifier les termes employés et de les définir de façon consensuelle et cohérente tout en préservant la diversité langagière.

La terminologie, en tant que discipline scientifique, se fonde sur une conceptualisation d'un domaine et sur les mots pour en parler. Elle se doit donc de concilier un point de vue linguistique et un point de vue ontologique. Elle doit également, dans une société numérique où les connaissances constituent la principale richesse, pouvoir être opérationnalisée à des fins de traitement de l'information.

Les conférences TOTh se situent dans le prolongement des colloques annuels de la Société française de terminologie organisés en décembre à Paris (Ecole normale supérieure de la rue d'Ulm). Planifiées à mi-parcours, au mois de juin à Annecy (Polytech'Savoie), elles en complètent l'offre et proposent des conférences avec appel à communications, comité de lecture et publication des actes.

Les conférences TOTh ont pour objectif de rassembler industriels, chercheurs, utilisateurs et formateurs dont les préoccupations relèvent à la fois de la terminologie et de l'ontologie et, de façon plus générale, de la langue et de l'ingénierie des connaissances. Elles se veulent un lieu d'échange et de partage où sont exposés problèmes, solutions et retours d'expériences tant sur le plan théorique qu'applicatif ; ainsi que les nouvelles tendances et perspectives des disciplines associées : terminologie, langues de spécialité, linguistique, intelligence artificielle, systèmes d'information, ingénierie collaborative, etc.

Christophe Roche, Président du Comité Scientifique

http://www.porphyre.org

Institut Porphyre
Savoir et Connaissance

ISBN 978-2-9536168-0-4
EAN 9782953616804

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1er juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy – 5 et 6 juin 2008

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2009. *Actes de la troisième conférence TOTh - Annecy – 4 & 5 juin 2009*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2009
ISBN 978-2-9536168-0-4
EAN 9782953616804

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2009

Annecy – 4 & 5 juin 2009

avec le soutien de :

- Société française de terminologie
- Association Européene de Terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

http://www.porphyre.org

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Pierre Blanc	EDF SEPTEN
Danièle Bourcier	CNRS, CERSA Paris
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille III
Viviane Cohen	France Télécom, Paris
Rute Costa	Professeur, Université Nouvelle de Lisbonne
Luc Damas	MCF, Université de Savoie
Sylvie Després	MCF, Université Paris XIII
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section de terminologie
Jean-Yves Gresser	ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Jean-Paul Haton	Professeur, Université de Nancy 1
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Université Paris XIII
Widad Mustafa	Professeur, Université de Lille III
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université du Québec en Outaouais
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

Dès la troisième édition, les conférences TOTh ont trouvé une structuration qui traduit bien à la fois le caractère scientifique et pluridisciplinaire de la terminologie et l'intérêt de notre communauté pour d'autres domaines partageant des préoccupations communes.

Ainsi, la conférence d'ouverture a été donnée par une personnalité invitée issue d'une discipline différente de la nôtre – ici la phylogénèse – mais pour laquelle le langage et la pensée jouent également un rôle primordial.

Les contributions se sont réparties naturellement, et par le jeu des évaluations de façon équitable, en trois groupes ayant donné lieu à trois sessions.

Le premier groupe a rassemblé les articles portant principalement sur la dimension linguistique de la terminologie. Ont été abordés l'extraction terminologique à partir de dictionnaire, la place accordée aux corpus dans la construction de terminologies, l'acquisition de connaissances à partir de textes et l'apport des ressources linguistiques issues du web.

La deuxième session s'est donc logiquement intéressée à la dimension conceptuelle de la terminologie. Les notions de concept, de relation, d'ontologie ont été au cœur des présentations portant sur les cartes conceptuelles pour les bibliothèques numériques, les relations dynamiques et les graphes conceptuels, l'alignement d'ontologies et l'accès multilingue aux ontologies.

Enfin, la troisième session a été consacrée à la présentation de plusieurs applications terminologiques pour des secteurs aussi différents que l'ingénierie nucléaire, l'informatique, le domaine bancaire ou l'agriculture biologique. Il est à souligner que ces applications ont permis d'aborder différents points théoriques tels que la variation terminologique, la diachronie ou la structure des dictionnaires.

La richesse des débats qui ont animé ces deux jours de conférence – chaque présentation, questions comprises, s'est vue allouer plus de quarante cinq minutes de temps de parole – a été certainement une des plus belles récompenses pour les participants de TOTh 2009.

Christophe Roche

Président du Comité Scientifique

Table des matières

SESSION 3

Approche lexico-sémantique de l'extraction terminologique : utilisation de ressources lexicographiques et validation sur corpus

Bertrand Gaiffe, Evelyne Jacquey, Laurence Kister

Résumé : Cet article présente une expérience d'extraction de terminologie à partir d'un dictionnaire en vue d'annoter des textes de spécialité par l'intermédiaire de leurs termes. Il décrit la méthode d'extraction de la terminologie et la méthode d'annotation des textes. Les difficultés liées à l'ambiguïté de forme de certains termes ("aspect" dans le domaine linguistique, par exemple) sont abordés ainsi que quelques solutions destinées y faire face : utilisation d'un dictionnaire (extraction de collocations) et de techniques endogènes habituelles pour l'extraction de candidats termes (patrons syntaxiques) avec l'utilisation de modifieurs recensés comme relevant du domaine considéré par le dictionnaire.

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

1. Introduction

L'augmentation et la diversification des échanges, la généralisation de l'Internet, conduisent à une explosion de la quantité d'informations textuelles à laquelle il est possible d'accéder très facilement. Chaque domaine scientifique est en constante évolution et on constate de plus en plus d'interpénétrations des domaines de spécialités du fait des nombreuses questions abordées de manières multi- ou interdisciplinaires. La difficulté relative à cette évolution réside dans la faible précision de l'information obtenue et son manque d'exhaustivité. Comme l'ont souligné (Bourigault & Aussenac-Gilles 2003), les ressources terminologiques et ontologiques, leur extraction et leur structuration constituent une contribution majeure pour les travaux touchant l'information et la documentation (Roche 2004), l'édition (El Mekki & Nazarenko 2002), la recherche d'informations, la classification de documents (Sanjuan & Ibekwe-Sanjuan 2002), ou encore la détection de documents à caractère raciste sur la Toile (Valette & Grabar 2004). Comme le soulignent la plupart des travaux sur le sujet, notamment (Bourigault *et al.* 2001) et (L'Homme 2004), l'extraction automatique et la structuration de ressources terminologiques (Nazarenko & Hamon 2002) ont déjà donné lieu à la création de nombreux outils s'appuyant sur différentes méthodes (symbolique/numérique), prenant en compte différents types d'informations pour la structuration (regroupements sur la base de relations hiérarchiques d'ordre conceptuel, de relations de sémantique lexicale, de convergences et de divergences entre termes, de l'analyse distributionnelle des contextes d'apparition des termes dans les textes, etc). Les travaux évoqués ci-dessus partagent un point méthodologique important : les ressources terminologiques produites sont essentiellement extraites à partir de textes liés à un domaine de spécialité, qu'il s'agisse d'une extraction automatique, manuelle ou assistée. L'utilisation de ressources externes, comme par exemple l'utilisation de dictionnaires, est seconde (L'Homme 2004).

Les travaux présentés ici proposent de partir d'une ressource lexicographique de référence fortement domanialisée, le Trésor de la Langue Française informatisé (TLFi) (Dendien & Pierrel 2002). Cette ressource comporte de nombreux sens explicitement associés à des domaines de spécialité (97330 sens domanialisés sur 271165, soit près de

36 %[1]). A partir d'une terminologie extraite automatiquement pour le domaine des sciences du langage, nous procédons à une validation sur un corpus spécialisé grâce à la détection automatique des termes dans les textes. Cette approche permet de contribuer à l'avancement des recherches dans trois domaines : l'information et la documentation, la linguistique textuelle et l'analyse des relations sémantiques et discursives.

Dans le domaine de l'information et de la documentation, plusieurs auteurs ont montré qu'il est possible d'indexer des textes à partir d'une terminologie ou d'un thésaurus (Bourigault *et al.* 2004) et (Aussenac-Gilles *et al.* 2000). Sur ce point particulier, nous avons montré dans des travaux antérieurs que Thésaulangue, le thésaurus constitué au laboratoire et intégré au portail terminologique de l'INIST, n'est pas totalement satisfaisant pour l'indexation et la classification des documents par les documentalistes (Kister & Jacquey 2007a et b ; Kister, Jacquey & Gaiffe 2008). Disposer d'une terminologie validée sur un corpus scientifique de spécialité constitue une plus-value pour le travail d'annotation automatique ou semi-automatique que nous envisageons.

Dans le domaine de la linguistique textuelle, la poursuite de nos travaux consacrés à la confrontation de la structure hiérarchique d'une terminologie et de la structure thématique des textes spécialisés correspondants est nécessaire. La structure thématique est appréhendée par repérage et étiquetage des termes dans les textes en tenant compte des différentes variations qu'ils subissent comme, par exemple, la reprise anaphorique. Ces travaux s'inscrivent dans la mise en regard de la terminologie et de la linguistique textuelle (Poibeau 2005) et dans le développement d'une terminologie textuelle (Bourrigault & Slodzian 1999).

L'importance du repérage et de l'étiquetage des termes dans les textes nous permet d'étudier 'in vivo' les relations sémantiques et discursives (L'Homme 2004) qui pourraient amender et améliorer la terminologie. L'analyse de ces relations doit permettre de mieux structurer la terminologie et par la même le thésaurus.

2. Extraction de la terminologie à partir du TLFi

La définition de "terme" que nous adoptons est fondée sur celle proposée par (L'Homme 2004) qui montre et synthétise l'évolution sémantique et conceptuelle de celui-ci depuis (Wüster 1981). A la suite de

1 Pour déterminer cette proportion, nous faisons l'hypothèse que chaque définition dans chaque bloc cohérent d'information du point de vue lexicographique représente un sens du lemme défini ou de l'élément de composition.

L'Homme, nous considérons un terme comme "une unité lexicale associée à un domaine de spécialité", pouvant être réalisé sous forme simple (lexème) ou complexe (syntagmes). Comme mentionné dans l'introduction, la plupart des travaux en extraction de terminologie utilisent des critères syntaxiques et des critères statistiques pour identifier des candidats termes. En ce qui nous concerne, les termes sont extraits à partir d'une ressource lexicographique. Nous nous sommes limités dans un premier temps à la catégorie du nom - les autres catégories grammaticales ne sont impliquées que dans la mesure où elles apparaissent dans des termes complexes (consonne fricative sonore, changement de sens) globalement considérés comme des noms complexes.

La ressource lexicographique que nous utilisons est le Trésor de la Langue Française informatisé (TLFi). Ce dictionnaire jouit d'une couverture remarquable pour le français des 19ème et 20ème siècles. Il comporte ~92 000 entrées (principales et secondaires) parmi lesquelles ~90 000 concernent des unités lexicales et grammaticales. Du point de vue lexical, il comporte :

- ~9 000 entrées de verbes
- ~16 500 entrées d'adjectifs
- ~48 300 entrées de noms
- ~11 700 entrées de noms et d'adjectifs confondus

ce qui correspond à ~85 500 entrées qui produisent ~278 500 définitions dont ~94 800 dépendent d'un domaine technique explicite, soit une proportion de ~34%. Dans le cadre d'expériences faisant intervenir une approche TAL, le dictionnaire est utilisé sous son format XML et étiqueté en catégories grammaticales pour les définitions et les exemples. Dans cette version XML, les techniques classiques de transformation XSLT permettent d'atteindre les objets lexicographiques balisés. Ceux qui nous intéressent ici, sont :

- la vedette <ved>
- le code grammatical <cod>
- les blocs d'information <B>
- les indicateurs d'emploi <ind>
- l'indication d'un domaine <dom>
- le texte de la définition <def>
- les conditions d'usage <cro> – délimitées par des crochets
- les synonymes et antonymes <syno>
- les syntagmes illustratifs <syntita n=i> qui sont les constructions courantes du lexème

- l'organisation hiérarchique <H>

Il faut encore préciser un point particulier concernant les collocations définies, c'est-à-dire les objets lexicographiques proches de ce qu'on appelle des expressions figées dans la littérature linguistique. Dans le dictionnaire, ces éléments sont considérés comme des entrées à part entière et sont repérables automatiquement car balisés par <syntita n=d>, directement suivi d'une définition <def>.

2.1. Procédure d'extraction

Dans le TLFi, nous avons sélectionné le domaine Sciences du Langage qui compte plusieurs sous domaines parmi lesquels nous avons sélectionné les sous domaines suivants[2] : grammaire, lexicographie, lexicologie, linguistique, philologie, phonétique, phonologie, rhétorique, sémiologie, sémiotique, stylistique, toponymie.

La procédure d'extraction consiste, dans un premier temps, en une feuille XSLT qui extrait du TLFi les informations pertinentes relatives à chaque occurrence d'un des domaines mentionnés. Si la vedette est de catégorie « substantif », la feuille extrait cette vedette, son code grammatical, les définitions relatives au domaine et les éventuels synonymes ou antonymes. Si, au contraire, la mention de domaine est dominée hiérarchiquement par un syntagme défini, on extrait sa définition et ses éventuels synonymes et antonymes. Dans un second temps, la terminologie extraite est revue manuellement pour calculer des variantes formelles. L'extraction de ce type de variantes fait l'objet de procédures automatisées sur l'italien (Dell'Ortella *et al.* 2008) pour des corpus juridiques et environnementaux. Outre la réduction de formes de vedettes telles "aberrant, -ante" ramenée à "aberrant", cette étape permet de ramener des termes complexes tels "Grammaire comparée ou linguistique comparée" aux variantes "Grammaire comparée" et "linguistique comparée"[3]. La dernière tâche réalisée à cette étape est l'étiquetage en parties du discours de chacune des variantes isolées.

2.2. Terminologie extraite pour les sciences du langage

La terminologie extraite à partir du TLFi comporte 2 402 entrées. En comparaison, le thésaurus initialement constitué dans le domaine des

2 Les sous domaines utilisés dans le TLFi sont accessibles dans la recherche assistée.

3 Dans le TLFi, parce qu'il a d'abord été édité sous la forme papier et pour limiter la place occupée par les informations, les lexicographes ont établi plusieurs procédés typographiques leur permettant de factoriser l'information. Dans la mesure où la terminologie devait être utilisée pour étiqueter des corpus, il a fallu expanser les informations factorisées.

sciences du langage au laboratoire comporte 872 entrées. Le nombre d'entrées est multiplié par 2,75. Cependant, l'augmentation en quantité ne se fait pas forcément à qualité constante, c'est pourquoi nous procédons ensuite à un étiquetage sur corpus et nous analysons quantitativement et qualitativement les résultats.

Du point de vue qualitatif, la terminologie, contrairement au thésaurus, n'est pas structurée mais elle est plus riche et plus précise car chaque entrée de la terminologie dispose de l'ensemble des informations lexicographiques extraites dans le bloc d'information correspondant à l'un des sous-domaines des sciences du langage. Ces informations lexicographiques concernent les conditions d'emploi, les liens de synonymie éventuels, les syntagmes illustratifs du sens domanialisé et un ou plusieurs exemples, le cas échéant. Dans l'exemple ci-dessous, le terme "voyelle" n'est pas seulement atteint, mais il est aussi associé aux informations présentes dans la ressource lexicographique qui fournit l'indication du domaine, la définition de "voyelle", un renvoi synonymique et une locution courante dans laquelle le terme apparaît.

```
<terme>
  <ved xml:id="e1876">voyelle</ved>
  <categorie>voyelle:n</categorie>
  <dom>PHONÉT.</dom>
  <def>Phonème constituant à lui seul un son ...</def>
  <syno>Synon. vx voix (v. ce mot I A 1)</syno>
  <syntagme_illustratif>Système des voyelles françaises.</syntagme_illustratif>
</terme>
```

En comparaison avec les approches d'extraction de termes à partir de textes, l'avantage est le fait de réduire la vérification manuelle :

- le dictionnaire - la ressource initiale - a été vérifié par des experts linguistes au moment de sa rédaction, il n'est pas nécessaire, à cette étape, de faire appel à des experts du domaine
- le nombre très restreint de termes extraits réduit grandement le coût des traitements manuels
- les variantes associées aux termes - au nombre de 472 - ont été vérifiées et munies des informations lexicographiques du bloc d'information extrait en fonction de l'étiquette de domaine

  A titre de comparaison, (Aussenac-Gilles & Bourigault 2000) extraient 21068 candidats de fréquence supérieure ou égale à 1 à partir de deux corpus :

- le corpus AFIA – riche de 31 212 occurrences - qui regroupe des descriptions de laboratoire
- le corpus LIVRIC – riche de 178 336 occurrences - qui contient des publications scientifiques

L'ensemble des termes candidats repérés, après avoir été filtrés en fonction du nombre de documents dans lesquels ils apparaissent et/ou leurs fréquences, sont ensuite évalués manuellement par les auteurs des articles. Chaque évaluateur évalue en moyenne 81 candidats extraits du corpus LIVRIC et 48 candidats repérés dans le corpus AFIA.

Pour la terminologie que nous avons constituée à partie du TLFi, le traitement manuel a consisté à vérifier les codes grammaticaux, l'expansion des alternatives, des énumérations et des optionalités ainsi que la qualité de l'extraction opérée automatiquement.

3. Repérage et étiquetage des termes en corpus

3.1. Le corpus

Le corpus de spécialité constitué pour l'expérience que nous présentons ici est d'une taille raisonnable selon (Bourigault & Aussenac-Gilles 2003) : il comporte 149 772 occurrences. Il est relativement homogène et cohérent avec la ressource lexicographique qui a permis d'extraire la terminologie que nous désirons valider sur ce corpus. Par la suite, nous envisageons d'appliquer la méthode à d'autres documents - le corpus des journaux du CNRS[4] - du domaine des sciences du langage ou d'autres domaines apparaissant dans le TLFi afin de partir d'une terminologie initiale de même nature que celle utilisée pour cette expérience.

Le corpus actuel compte trois œuvres fondatrices en linguistique datant du 20ème siècle : Cours de linguistique générale de Ferdinand de Saussure [1916], Le langage et la vie de Charles Bally (1952) et La linguistique de Jean Perrot (édition de 1989).

3.2. Procédure de détection et d'annotation

Pour repérer les termes dans les textes, nous avons étiqueté les textes en parties du discours en utilisant TreeTagger, nous avons ensuite identifié

4 Des négociations, à l'initiative du service de communication et de valorisation de la recherche du laboratoire sont en cours afin que nous puissions utiliser ces documents sous droit.

les termes grâce aux variantes également étiquetées en parties du discours. L'étiquetage morpho-syntaxique est évidemment indispensable pour s'affranchir de confusions de formes telles : "son" adjectif possessif versus "son" nom commun et terme dans le domaine qui nous intéresse. Outre cet exemple anecdotique, l'étiquetage en morpho-syntaxe permet d'éviter un grand nombre de confusions nom commun/adjectif.

Nous avons étiquetés tous les termes, même si nous ne retenons in fine que ceux d'extension maximale :

<terme ref='e630'><terme ref='e637'>grammaire</terme>comparée</terme>

Notons enfin que chaque occurrence de terme annotée fait référence (via l'attribut <ref>) à la terminologie extraite, ce qui permet, si nécessaire, de se reporter aux informations lexicographiques associées au terme.

3.3. Résultats quantitatifs

Comme le montrent les décomptes ci-dessous (Tableau 1 - Densité en termes par rapport aux noms présents dans le corpus), le nombre d'occurrences de termes reconnus parmi les noms présents dans le corpus est multiplié par 10 par rapport au nombre d'entités dans le thésaurus : 3% des noms sont étiquetés comme des termes à partir de la nomenclature du thésaurus quand 32% des noms sont étiquetés comme des termes à partir de la nomenclature complète de la terminologie, termes et variantes. Parallèlement, le nombre de termes différents reconnus passe de 132 avec le thésaurus à 451 avec la terminologie.

	Corpus	Nb Noms	Thésaurus	Terminologie
Occurrences	149 772	54 119	1716 - 3%	17374 - 32%
Candidats Termes	5 662	2 629	132	451

Tableau 1. Densité en termes par rapport aux noms présents dans le corpus

L'exemple ci-dessous montre un résultat d'annotation (termes reconnus en gras et expressions référentielles en fonction de l'étiquetage[5], soulignées).

La numération intéresse tous les **aspects** du langage : de la **phonologie** (**nombre** et **fréquence** des **phonèmes** dans une **langue** donnée) à la **syntaxe** (par exemple , **fréquence relative** des différentes dispositions possibles dans la **phrase** pour les éléments constituants) , au **lexique** (des dénombrements en montrent l'**extension** , liée aux besoins auxquels le **vocabulaire** doit répondre) et à la **stylistique** , qui tend de plus en plus à prendre pour **base** , dans l'appréciation des faits individuels, des statistiques de **fréquence** des différentes **réalisations** possibles dans la **langue** .

4. Analyse des résultats obtenus

Un premier regard sur les résultats d'étiquetage permet de constater que la procédure donne de bons résultats tant que les formes des termes ne sont pas ambiguës et qu'il s'agit, lorsqu'ils sont réalisés de différentes manières, de leur réalisation maximale. Cependant, les termes peuvent apparaître sous une forme ambiguë, soit par nature, soit du fait de l'élision d'éléments par rapport à sa forme maximale. Ainsi « objet », par exemple, a une définition terminologique très claire, il s'agit du complément d'objet direct, mais ce nom a aussi un sens très général qui est celui que l'on trouve dans "cette étude a pour objet". De la même manière, "aspect" est reconnu à tort comme terme dans l'exemple ci-dessus dans "tous les aspects du langage" alors qu'il a un sens non équivoque en sciences du langage quand on s'intéresse à la conjugaison.

Pour aller au delà de cette première intuition, nous comparons les fréquences relatives des termes présents dans le corpus de spécialité ~150 000 occurrences - données à 1 pour 1 000 - avec celles que l'on observe dans un corpus ne relevant pas de cette spécialité constitué de deux journées complètes de l'Est Républicain toutes éditions locales confondues soit ~682 000 occurrences. Bien entendu, les deux corpus reçoivent les mêmes prétraitements (normalisation, étiquetage morpho-syntaxique, lemmatisation). La comparaison des fréquences est illustrée sur le graphique suivant pour les formes les plus fréquemment rencontrées dans le corpus de spécialité.

5 Les étiquettes utilisées sont du type NP pour nom propre, PR pour pronom, etc.

	Langue	forme	son	cas	rapport	temps	analogie	sujet	objet	lieu
F1[6]	0,76	0,25	0,16	0,14	0,13	0,10	0,06	0,05	0,07	0,04
F2[7]	0,04	0,10	0,04	0,21	0,10	0,60	0	0,08	0,15	0,75
D	0,72	0,15	0,12	0,07	0,03	0,50	0,06	0,03	0,08	0,71

Tableau 2. **Fréquences relatives sur le corpus de spécialité et le corpus tout venant**

Au vu de ce tableau, deux cas de figures apparaissent. De façon très claire, un terme comme "langue" est peu ambigu dans le corpus de spécialité bien que très fréquent dans la mesure où il est au contraire peu fréquent dans le corpus tout venant. Les formes "lieu" et "temps" ont un comportement exactement inverse. Il s'agit probablement d'éléments qui différencient les deux types de corpus et on peut donc penser que leurs emplois dans le corpus de spécialité sont essentiellement terminologiques. Le second cas de figure est illustré par des formes telles "son", "cas", "objet" ou "sujet" pour lesquelles on peut soupçonner que l'étiquetage en termes fondé sur leur seule forme demande à être vérifié. Deux pistes sont actuellement explorées :

- éliminer de l'analyse des emplois relevant de collocations : "au sujet de", "dans tous les cas", etc.
- enrichir la terminologie de termes complexes construits à partir de ces termes de forme ambiguë : "cas nominatif", "sujet profond", etc.

4.1. Détection des collocations et filtrage

Deux sources fournissent des collocations. La première consiste à reprendre l'ensemble des collocations référencées dans le TLFi si elles ne sont pas associées au domaine considéré. Les collocations suivantes ont été trouvées pour "sujet" : "sujet (battu et) rebattu, à ce sujet, au sujet de, sujet psychologique, sujet-contact, sujet de la connaissance, sujet transcendantal, sujet secondaire". La seconde source est fournie par tout corpus tout venant dont, une fois la liste de formes à vérifier connue, il est aisé d'extraire les collocations les plus fréquentes.

En pratique, il faut croiser les deux sources : la notion de collocations "les plus fréquentes" dans un corpus suppose le choix d'un seuil difficile à

6 Fréquence 1 = fréquence relative dans le corpus de spécialité
7 Fréquence 2 = fréquence relative dans le corpus 'tout venant'

fixer. Par ailleurs, les collocations du TLFi ne peuvent pas être exploitées directement, comme pour la terminologie extraite, elles demandent quelques corrections manuelles destinées à expanser et vérifier l'information transcrite de manière économique dans le dictionnaire.

4.2. Enrichissement en termes complexes

Pour extraire la terminologie, nous avons pris le parti de ne considérer que les substantifs étant donné le rôle central joué par cette catégorie comme le souligne la littérature sur le sujet. Comme le mentionne (L'Homme 2004), les adjectifs et les verbes participent à la définition d'une terminologie exhaustive. C'est également la stratégie à l'œuvre dans LEXTER qui considère des patrons dont l'expression maximale est ADJ? NOM [NOM | ADJ | de]*. Pour détecter des candidats termes supplémentaires (qui permettront de désambiguïser certains des emplois des termes de forme ambiguë), nous avons extrait les adjectifs relatifs au domaine des Sciences du Langage et nous avons récolté les "chunk" auxquels ces adjectifs participent. On récolte de cette façon 213 candidats termes dont 56 sont des termes nouveaux. Parmi eux, certains permettent de désambiguïser, par exemple, "rapport" dans "rapports syntagmatiques".

L'inconvénient, en revanche, de ces méthodes est que les candidats termes ainsi récoltés n'ont pas de définitions.

5. Conclusions et perspectives

Dans cet article, nous avons explicité une procédure permettant d'extraire une terminologie des Sciences du Langage à partir de la ressource lexicographique que constitue le TLFi. Dans sa forme XML, le balisage des objets lexicographiques tels la vedette, le domaine d'emploi et la définition permettent l'extraction, minimalement des triplets <ved>, <dom>, <def>. Un traitement manuel a ensuite été appliqué afin d'expanser les informations factorisées, notamment les différentes variantes possibles de collocations définies comme, par exemple, "complément d'objet direct", "complément d'objet", etc. Le second aspect du traitement manuel de la terminologie a consisté à étiqueter grammaticalement les termes pour lesquels les codes grammaticaux ne pouvaient pas être inférés à partir de l'objet lexicographique <cod>.

Afin de procéder à la validation de la terminologie extraite, et parce que nous nous inscrivons dans la perspective de deux types de recherches en cours :

- l'étude de l'évolution thématique des termes dans les textes, dans la philosophie de travaux du même ordre (Ferret & Grau 2006) et à la suite de travaux personnels antérieurs (Kister & Jacquey 2007a et b ; Kister, Jacquey & Gaiffe, 2008)

- l'étude des relations lexicales sémantiques entre termes

nous avons effectué l'étiquetage d'un corpus homogène, de taille raisonnable dans le domaine des Sciences du Langage.

L'analyse des résultats montre une nette amélioration quantitative par rapport au thésaurus utilisé initialement dans les perspectives de recherche évoquées ci-dessus. Sur le plan qualitatif, nous avons avancé plusieurs pistes de solution afin de traiter notamment la question de l'ambiguïté des formes des termes, même lorsqu'elles apparaissent dans un corpus de spécialité et qu'elles font référence à des termes non ambigus de la terminologie.

Pour la suite, nous envisageons de structurer la terminologie en prenant appui sur les travaux déjà menés dans cette perspective (Nazarenko et Hamon 2002). Deux axes seront mis en œuvre :

- un axe en domaines, autrement dit, structurer les termes selon l'organisation hiérarchique du thésaurus

- un axe spécifique/générique en exploitant les définitions associées aux termes

Pour ce qui est de l'étiquetage dans les textes, les solutions esquissées pour le traitement de l'ambiguïté des formes des termes seront intégrées dans la procédure automatique d'ensemble.

Bibliographie

Aussenac-Gilles Nathalie & Bourigault Didier (2000) : The Th[IC]2 Initiative : Corpus-Based Thesaurus Construction for Indexing WWW Documents, in Proceedings of the EKAW'2000 workshop "Ontologies and texts", Juan-Les-Pins, Université Paul Sabatier, Toulouse, pp. 71-78, octobre 2000

Bourigault Didier et Slodzian Monique (1999) : Pour une terminologie textuelle, in Terminologies nouvelles, 19, pp. 29-32

Bourigault Didier, Jacquemin Christian et L'Homme Marie-Claude (2001) : Recent Advances in Computational Terminology, Amsterdam/Philadelphie : John Benjamins

Bourigault Didier et Aussenac-Gilles Nathalie (2003) : Construction d'ontologies à partir de textes, Dans : Actes de la conférence TALN 2003, Batz-sur-Mer

Bourigault Didier, Aussenac-Gilles Nathalie & Charlet Jean (2004) : Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, Dans : Revue d'Intelligence Artificielle (RIA), Numéro spécial sur les techniques informatiques de structuration de terminologies, M. Slodzian (Ed.), Hermès, Paris, Vol. 18, N. 1/2004, pp. 87-110

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, 5-6-7 juin, AIDAinformazioni, 26, 1-2, pp. 185-206

Dendien Jacques et Pierrel Jean-Marie (2002) : Le trésor de la langue informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence, in TAL

El Mekki Touria et Nazarenko Adeline (2002) : Comment aider un auteur à construire l'index d'ouvrage, in Actes de la conférence CIFT 2002. pp. 141 – 157, Tunisie

Kister Laurence & Jacquey Evelyne (2007) : 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 Laurence & Jacquey Evelyne (2007) : 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)

Kister Laurence, Jacquey Evelyne & Gaiffe Bertrand (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, 5-6-7 juin, AIDAinformazioni, 26, 1-2, pp. 25-36

L'Homme Marie-Claude (2004) : La terminologie : Principes et Techniques. Presses Universitaires de Montréal

Nazarenko Adeline & Hamon Thierry (2002) : Structuration de terminologie, Editeurs de ce numéro dans TAL, vol 43, n°1, 174 p

Poibeau Thierry (2005) : Parcours interprétatifs et terminologie, in Actes de la conférence TIA 2005, Rouen

Roche Mathieu (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 Eric & Ibekwe-Sanjuan Fidelia (2002) : Terminologie et classification automatique des textes, in Actes de la conférence JADT 2002, pp. 677-688.

Valette Mathieu & Grabar Natalia (2004) : Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP, in Actes de la conférence JADT 2004

Wüster Eugen (1981) : L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses, in Textes choisis de terminologie 1, Fondements théoriques de la terminologie, Presses de l'université de Laval, pp. 55-114

A propos des auteurs

Bertrand Gaiffe

Lexique ATILF UMR 7118
44, avenue de la Libération BP 30687
54063 Nancy Cedex
www.atilf.fr
Bertand.Gaiffe@atilf.fr

Evelyne Jacquey

Lexique ATILF UMR 7118
44, avenue de la Libération BP 30687
54063 Nancy Cedex
www.atilf.fr
Evelyne.Jacquey@atilf.fr

Laurence Kister

Lexique ATILF UMR 7118
44, avenue de la Libération BP 30687
54063 Nancy Cedex
www.atilf.fr
Laurence.Kister@atilf.fr