Prise en compte de l'application dans la constitution de produits terminologiques

Nathalie Aussenac-Gilles[1], Anne Condamines[2], Sylvie Szulman[3]

(1) IRIT – CNRS, UPS, Toulouse, aussenac@irit.fr
(2) ERSS – CNRS, Maison de la recherche, Toulouse, anne.condamines@univ-tlse2.fr
(3) LIPN – CNRS, Univ. Paris Nord, Villetaneuse, ss@lipn.univ-paris13.fr

Résumé. Les produits terminologiques, de plus en plus sur support informatique, se trouvent utilisés dans différents types d'applications où textes et connaissances jouent un rôle privilégié. Leur constitution à partir de textes requiert de définir un cadre méthodologique situant l'usage d'outils de traitement de la langue. Nous montrons que la nature de l'application visée conditionne chacune des étapes de ce processus.

1 INTRODUCTION (1 PAGE)

Plusieurs phénomènes – comme la généralisation des documents électroniques au sein des entreprises, l'explosion d'Internet ou encore la volonté de réutiliser des expériences et des connaissances - conduisent à un développement croissant d'applications informatiques capables de gérer des connaissances, d'accéder à l'information contenue dans des textes, et de les restituer de la manière la plus pertinente possible aux utilisateurs. Plusieurs disciplines sont concernées par la mise au point de ce type d'application, qui vont des sciences de l'information et la terminologie à différentes composantes de l'informatique, comme la recherche d'information, le traitement du langage naturel et l'ingénierie des connaissances. La linguistique de corpus est également concernée puisqu'il s'agit de donner du sens à des unités textuelles.

Au cœur de la mise au point de ces applications se trouvent deux composantes directement liées à la langue : les corpus, qui sont les sources des connaissances ou les fonds à explorer pour l'application ; les structures, ressources ou produits terminologiques, qui sont des représentations informatiques des connaissances associées à la langue permettant à l'application de répondre aux besoins. Dans notre perspective, ces ressources terminologiques (thésaurus, terminologies, ontologies, …) sont construites à partir de textes et utilisées ensuite pour accéder à des connaissances dans ces textes ou d'autres documents. Elles jouent donc un rôle pivot et leur qualité est déterminante pour garantir l'usage de l'application.

La recherche de l'efficacité et de la rationalisation de la production des logiciels incite à constituer des ressources terminologiques les plus génériques et réutilisables possibles. Elle tend également à imaginer des outils génériques pour les construire à partir de textes, des logiciels de traitement de la langue basés sur des principes universaux qui seraient applicables sur tout corpus. Or la pratique d'une part et certaines bases théoriques linguistiques sur le lien entre le sens des mots et leur usage montrent que cette généricité va à l'encontre de la bonne adéquation à un usage particulier. Les travaux des membres du groupe TIA (Terminologie et Intelligence Artificielle) convergent pour étayer ce point de vue. Finalement, l'objectif final d'usage de la ressource terminologique a de multiples conséquences sur l'ensemble du processus qui va des corpus d'origine à l'application en passant par la constitution d'un produit terminologique. C'est ce que nous montrons dans cet article.

Dans une première partie, nous définissons les applications auxquelles nous faisons référence et la notion de ressource terminologique. Dans une deuxième partie, nous présentons précisément la problématique qui nous intéresse et mettons l'accent sur les méthodes et outils, principalement de traitement automatique de la langue et de modélisation des connaissances, utilisés dans ce processus. Enfin, dans une troisième partie, nous illustrons par trois exemples les impacts de l'application finale sur le développement d'une application : impact sur les méthodes et principes linguistiques d'analyse des textes, impact sur la représentation des connaissances et le choix des logiciels d'exploration de textes, impact sur le contenu même des ressources terminologiques.

2 DEFINITIONS

2.1 Ressources et produits terminologiques (Sylvie)

Par « produit terminologique », on entend un ensemble plus ou moins structuré de termes. Dans cet article, nous considérons que cet ensemble est obtenu comme résultat d'une analyse de corpus. Celle-ci peut être effectuée selon un processus assisté par des outils de TALN, mais le produit final est toujours validé manuellement. Lorsqu'il sert à créer une application, le produit terminologique s'appelle « ressource terminologique ».

La distinction entre les différents produits terminologiques se fait principalement sur la structure utilisée.

Produit	Structure	Statut	Contenu	Domaine couvert
lexique	liste alphabétique	non formel	mots	domaine (s)
Index	liste alphabétique - sous liste	non formel	mots - sujets - noms	livre - ouvrage
Glossaire	liste alphabétique	non formel	termes + définition en LN de termes d'une activité- d'un domaine	une activité - un métier
Terminologie	liste alphabétique ou réseau	non formel	termes + propriétés linguistiques + propriétés grammaticales + relations entre termes	une science - un art - un domaine
Thesaurus	réseau	formel	notions + relations prédéfinies	ensemble de domaines
Ontologie	réseau	formel	concepts organisés en taxinomie + relations entre concepts	un domaine tâches

Tableau 1 *: Les différents produits terminologiques*

2.2 Exemples d'applications

Les applications qui peuvent mettre en œuvre des ressources terminologiques sont diverses et peuvent être rattachées à différents domaines de recherche. Nous citons ci-dessous quelques exemples d'applications dans les domaines suivants :

- Traitement automatique des langues (TALN) : résumés automatiques (utilisation de lexiques et de terminologies [Royauté]), système de questions/réponses, aide à la traduction (lexiques bilingue ou ontologies) [Slodzian], compréhension de textes (lexiques ou ontologies)
- Recherche d'information (RI) : gestion de documents (classification, indexation et catégorisation) [Nazarenko], recherche d'information multi-média ([Dieng], [Bachimont]), extraction d'information et de

connaissances à partir de textes ([Nazarenko][Cherfi,Toussaint, 2002]

- Ingénierie des connaissances (IC) : veille technologique, mémoire d'entreprise [Dieng, Golebiowska], gestion des connaissances de l'entreprise [Cerbah] [Charlet]

Ces listes sont loin d'être exhaustives. Chaque type d'application utilise une ressource terminologique dont le contenu peut aussi varier en fonction de l'usage de l'application finale. C'est cette influence de l'application sur les tâches situées en amont du processus de constitution d'un logiciel en IC qui nous intéresse particulièrement dans cet article.

3 PROBLEMATIQUE

3.1 D'un corpus à une application

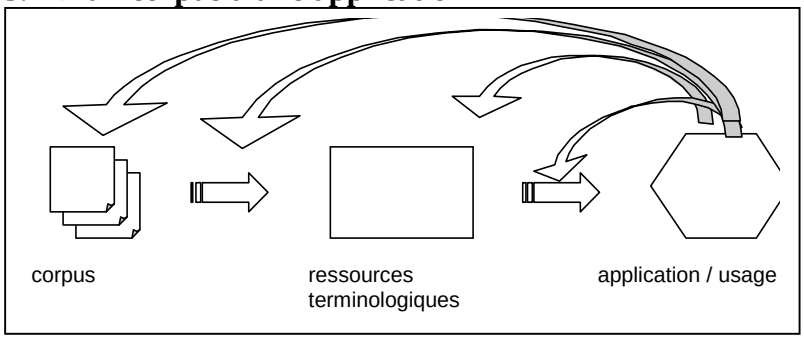


FIG. 1 – *Problématique du passage d'un corpus à une application*

Notre problématique est schématisée dans la figure 1. Elle intéresse plusieurs disciplines : l'informatique pour le développement d'outils d'aide (TALN) et d'applications (IC), mais aussi les sciences de l'information, qui fournissent des contextes d'usage, ou encore la terminologie, la lexicologie et la linguistique de corpus, autant pour l'étude de la nature des ressources terminologiques que la proposition de cadres théoriques et de méthodes pour accéder à des connaissances à partir de textes.

Quel que soit le point de vue disciplinaire duquel on se place, les ressources terminologiques jouent le rôle de médiateur entre des utilisateurs dans un contexte d'application donné, et des connaissances dont l'expression exhaustive se trouve dans les textes. Ces ressources ne

peuvent être construites de manière neutre et générique, indépendamment de la prise en compte de ce contexte. Sinon, elles risquent de n'être ni pertinentes ni utilisables.

Cette influence de l'application visée sur la construction de l'application et des ressources terminologiques se répercute donc autant sur le choix des textes sources formant le corpus que sur le choix des méthodes et outils les mieux adaptés à ce corpus et à la ressource à produire. Ces diverses influences sont matérialisées par les flèches sur la figure 1. Les paragraphes 4.1, 4.2 et 4.3 illustrent chacun certains de ces impacts. Auparavant, nous rappelons quels types de méthodes et outils peuvent être utilisés en fonction des objectifs. Nous étudions également la notion de genre textuel, qui nous semble tout à fait pertinente pour caractériser les corpus et juger de leur adéquation dans un projet donné.

3.2 Outils et méthodes

Il existe une panoplie d'outils pour chaque type de ressource et pour différents types de traitements sur les textes. Ces outils ont été construits en référence à des théories spécifiques, qui induisent des contraintes d'utilisation et des biais dans les résultats obtenus. Il est important de connaître ces présupposés avant de les choisir pour un projet donné. Leur utilisation sera donc plus ou moins pertinente en fonction de la nature du corpus et du produit terminologique à construire. Parmi les logiciels disponibles, nous distinguons trois types qui peuvent être utilisés pour différents types de ressources :

- Outils de TALN indépendants : SynoTerm [Nazarenko] exploite la compositionnalité des mots pour proposer des synonymes de termes complexes, Syntex [Bourigault] focalise la recherche de termes sur des mots régulèrement en dépendance syntaxique, Caméléon [Aussenac Séguéla, 2000] vise le repérage de relations lexicales à partir de régularités morpho-syntaxiques, analyses statistiques [Toussaint], analyseurs morphologiques [Zweigenbaum], Astrium [Faure Nedellec]
- Plates-formes intégrant des traitements de la langue : LIKES [Rousselot], TERMINAE [Biébow Szulman, 2002] pour la construction de terminologies et d'ontologies
- Plates-formes de modélisation : DOE [Bachimont, Troncy], OILEdit, Protégé2000, Likes [Rousselot]

Les méthodes (comme DOE [Bachimont, Troncy], OntoSpec [Kassel], TERMINAE [Biébow Szulman Aussenac]) cherchent à définir les modalités d'utilisation des outils et d'exploitation de leurs résultats. Elles

s'appuient sur un langage de représentation des connaissances permettant de décrire le modèle. Le paragraphe 4.2 illustre comment mettre au point une méthode adaptée à une classe particulière d'applications.

3.3 Genres textuels (Anne)

La notion de genre textuel permet, dans une première approximation, de prendre en compte la situation de production d'un texte dans son analyse. En effet, un texte n'est pas seulement une manifestation langagière, c'est une manifestation langagière réalisée dans un certain contexte. Utilisée depuis longtemps dans le domaine littéraire où l'on parle du roman, de la nouvelle, de l'essai … comme d'autant de genres, cette notion est maintenant utilisée pour tous les types de discours. C'est Bakhtine qui, dès les années 1930, a essayé de théoriser la notion de genre discursif :

« chaque sphère d'utilisation de la langue élabore ses types relativement stables d'énoncés, et c'est ce que nous appelons la notion de genre » (Bakhtine,1984,265).

En d'autres termes, les caractéristiques extra-linguistiques sont corrélées avec des caractéristiques linguistiques repérables dans leurs régularités. La difficulté vient de ce qu'il est très difficile de caractériser *a priori* toutes les situations extra-linguistiques possibles :

« La richesse et la variété des genres du discours sont infinies car la variété virtuelle de l'activité humaine est inépuisable et chaque sphère de cette activité comporte un répertoire de genres du discours qui va se différentiant et s'amplifiant à mesure que se développe et se complexifie la sphère donnée ».(ibid.,265).

Des tentatives de classement des genres discursifs de textes ont été faites, comme dans le projet Eagles (http :www.ilc.pi.cnr.it/EAGLES96/textyp/texttyp.html)) ou dans la TEI (Text Encoding Initiative). Mais les classes obtenues restent très générales et souvent inadaptées aux besoins réels d'application. En revanche, pour des besoins en ingénierie des connaissances, qui concernent des textes issus du seul milieu de l'entreprise, la possibilité d'un classement semble envisageable. En plus des éléments habituellement utilisés pour caractériser les textes : date, statut des auteurs et des lecteurs, objectif du texte, … , un autre type d'élément pourrait être pris en compte qui concernerait l'utilisation possible de ce texte à d'autres fins que l'objectif initialement prévu. En effet, il faut noter que dans tous les cas, en IC, les ontologues utilisent des textes dont ils n'étaient pas les destinataires et pour une visée qui n'était pas intentionnelle de la part des rédacteurs

(construire des ressources terminologiques de différentes natures). L'expérience montre par ailleurs que le choix des textes pour constituer le corpus est dépendant de l'application : on ne sélectionne pas les mêmes textes selon que l'on souhaite construire une ressource terminologique pour faire de la recherche d'information ou pour faire un modèle du domaine.

Or, tous les textes ne sont pas également utilisables pour construire des ressources terminologiques : certains sont peu utilisables (par exemple s'ils sont d'un niveau d'expertise trop élevé et difficilement accessible à un néophyte), certains sont mieux adaptés pour construire une ontologie (ceux dont le genre didactique est manifeste par exemple) et d'autres pour construire un thésaurus (ceux qui contiennent des méta-connaissances). Si l'enjeu de l'IC devient suffisamment important dans les entreprises on pourrait envisager que la possibilité de prise en compte pour construire telle ou telle ressource terminologique constitue une caractéristique des textes, au même titre que le niveau d'expertise de l'auteur par exemple. Cette éventualité ne peut être validée que si l'on arrive à corréler la notion de construction d'une ressource terminologique particulière avec des fonctionnements linguistiques réguliers (cf 4.1).

4 ILLUSTRATIONS

Pour le moment, les travaux effectués ne nous donnent qu'une idée intuitive des incidences de l'application sur les différents paramètres de ce processus. Nos affirmations s'appuient sur des retours d'expérience qui n'ont pas encore été validés sur de nombreux corpus. Nous rendons compte ici de quelques projets qui illustrent notre propos.

4.1 Impact de l'objectif sur l'analyse de textes (Anne)

Une des méthodes linguistiques de repérage de relations conceptuelles se base sur la notion de « marqueurs de relation ». Il s'agit d'éléments linguistiques que l'on peut utiliser systématiquement pour donner une interprétation relationnelle à certains contextes. Par exemples : `tous les N1 sauf dét N2` (*tous les mammifères sauf la baleine*) peut permettre de décider qu'il y a une relation de spécifique à générique entre N2 (*baleine*) et N1 (*mammifère*). Or, justement pour ce qui concerne la relation de généricité, il semblerait que tous les marqueurs ne donnent pas le même type de résultats. Ainsi, le marqueur qui consiste en une reprise anaphorique d' un nom par un autre nom qui est son *générique (La baleine s'approchait. Ce mammifère…*) semble permettre de repérer des génériques de plus haut niveau dans la hiérarchie que des marqueurs

plus classiques de généricité comme : `dét N1 être dét N2 + différences`.
Une étude a été réalisée sur un corpus d'environ 350 pages sur le génie logiciel (corpus fourni par EDF).
Le marqueur anaphorique a permis de repérer 44 génériques (ex : *Archivage de l'état de Configuration Logiciel. Cette activité est à la charge du responsable de la gestion de configuration*).
Sur ces 44 génériques, seuls 12 sont des termes complexes, ce qui est particulièrement étonnant quand on sait que la majorité des termes sont des combinaisons nominales. Par ailleurs, aucun de ces termes ne se trouve ni en position de N1 ni en position de N2 dans la structure « classique » proposée ci-dessus. Or, cette structure est la structure définitoire la plus anciennement décrite (par Aristote). Cela signifie donc que les génériques trouvés grâce à la structure avec anaphorique ne sont jamais utilisés pour définir un autre terme mais ils ne sont pas non plus définis dans le corpus. En revanche, ces mêmes génériques apparaissent comme têtes de termes complexes :
Acteur (tête d'un terme), *activité* (tête de deux termes), *composant* (tête de deux termes), *décomposition* (tête de deux termes), *espace* (tête de trois termes), *phase* (tête de deux termes), *processus* (tête de deux termes), *revue* (tête d'un terme).
Ainsi, il est clair que, d'un point de vue linguistique, ces termes ont un fonctionnement original ; quant à leur sens, il semble moins spécialisé que celui des termes trouvés avec d'autres marqueurs. Ces termes pourraient donc peut-être être utilisés pour constituer des ressources terminologiques qui ne s'adressent pas à des experts mais à des non-spécialistes.
Par ailleurs, si certains textes sont plus riches en marqueurs du type anaphorique, alors, ce pourrait être une façon de sélectionner les textes pertinents pour constituer une ressource « grand public », alors que les textes riches en marqueurs « classiques » seraient utilisés pour la constitution de ressources spécialisées. Cela reste une hypothèse qu'il faudra vérifier mais il reste très réel que les différents marqueurs ne permettent pas d'obtenir les mêmes types de termes et que cette possibilité mériterait d'être d'abord examinée de manière plus approfondie et ensuite éventuellement utilisée de manière régulière.

4.2 Impact de l'objectif sur le choix de la méthode de construction et le choix de la structure terminologique (Sylvie d'abord puis Nathalie)

Quel index pour le document électronique ? travaux de thèse de Touria Aït El Mekki. L'application = consulter un document sur support électronique. Le 1er impact (de l'application sur la structure) = le choix de la structure terminologique adaptée qui va constituer l'index. La structure classique de l'index papier ne convient plus et doit être enrichie pour devenir une des structures possibles des produits terminologiques. Par exemple, il est intéressant de disposer de liens sémantiques entre les entrées de l'index plus riches que les seuls liens *voir aussi* ou le lien de spécialisation.

le 2° impact (de la structure sur la méthode et les outils) : quelle est la bonne méthode et quels les bons outils pour construire cet index ? La méthode choisie combine l'utilisation de plusieurs outils de TAL et des techniques d'apprentissage de manière à s'adapter à plusieurs genres textuels.

4.3 Impact de l'objectif sur le contenu de la structure terminologique

Bien que l'on cherche le plus souvent à produire une vue consensuelle du domaine au sein d'une structure terminologique (ceci est particulièrement souhaité dans une ontologie), nous défendons l'idée qu'une ressource terminologique ne peut être construire sans que l'on sache à quoi elle va servir. En effet, l'organisation de termes et de concepts au sein d'un réseau requiert un ensemble de choix. A partir d'une même source (ici, des textes), plusieurs personnes vont produire des modèles différents parce qu'elles se font une représentation différente de ces connaissances et surtout de leur utilisation. Afin de converger vers un contenu adapté à l'application visée, il est donc important d'expliciter ces choix et de les justifier également en fonction de l'application. Ces choix concernent toutes les décisions de représentation, comme les exemples ci-dessous liés à la construction d'une ontologie :

- décision ou non de définir un concept ou une relation à partir d'une information dans les textes : certains passages peuvent être ignorés car jugés inutiles pour l'application ;
- place d'un concept dans la hiérarchie : dans une ontologie du tourisme, la distinction entre « moyen de transport international » et « moyen de transport urbain » conduit à classer « voiture » sous le premier concept

et « voiture de location » sous le deuxième concept. En fait, on différencie ici les concepts en fonction de leur utilisation et non de ce qu'ils sont physiquement, ce qui aurait conduit à considérer « voiture de location » comme un concept fils de « voiture » (EKAW2002)

- organisation globale des concepts : pour représenter des logiciels d'ingénierie des connaissances (projetTh(IC)2), plusieurs structurations des concepts de haut niveau ont été envisagées : l'une couvrait tous les concepts de l'ingénierie des connaissances (méthodes, outils, langages), parmi lesquels se situent les outils ; l'autre se focalise sur les fonctionnalités des outils (aide à la modélisation, aide à l'acquisition, aide à la validation). Le premier choix semble mieux adapté à une présentation de travaux de recherches par des équipes, l'autre guiderait mieux un cogniticien à la recherche d'un outil pour un objectif donné. Ces deux visions correspondent donc à deux applications différentes.

5 CONCLUSION (VENDREDI)

Lien avec ASSTICCOT et lien avec la RI

Quels objectifs de recherche ? quels outils ? pb de l'évaluation

Côté explosif car croisement des influences et des paramètres

6 REFERENCES

Les références sont numérotées par ordre alphabétique.

[1] N. Aussenac-Gilles, B. Biébow and S. Szulman, Revisiting Ontology Design : a methodology based on corpus analysis, Knowledge Engineering and Knowledge Management : Methods, Models, and Tools. Proc. of the 12th International Conference, (EKAW'2000), Springer-Verlag, Ed. R. Dieng and O. Corby, LNAI 1937, pp. 172-188,

[] AUSSENAC N., SEGUELA P., Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire*, N° spécial sur la linguistique de corpus. A. Condamines (Ed.) Vol 25. Déc. 2000. Toulouse : Publication de l'ERSS. Pp 175-198.

Bakhtine M. [1984] Esthétique de la création verbale. Paris : Tell,Gallimard.

Condamines A. [2002] : Corpus Analysis and Conceptual relation Patterns. *Terminology. 141-162.*

FAURE D. & NEDELLEC C. [1999] Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM Proc. of the *11th European Workshop, Knowledge Acquisition, Modelling and Management (EKAW'99)*, Juan-les-Pins, France, 329-334.

ROUSSELOT F., URL de LIKES http://www-ensais.u-strasbg/LIIA/likes/likes.html

[2] SZULMAN S., BIEBOW B. & AUSSENAC-GILLES N. (2002), Structuration de Terminologies à l'aide d'outils d'analyse de textes avec TERMINAE, *TAL*, Paris : Hermès. à paraître.