Représentation sémantisée des textes : terminologies et dimensions pragmatiques (qui-quand-où)

Nathalie Aussenac-Gilles

Représentation sémantisée des textes :

terminologies et dimensions pragmatiques (qui-quand-où)

N. Aussenac-Gilles

IRIT – UPS, 118, route de Narbonne – F-31062 Toulouse Cedex

aussenac@irit.fr

RÉSUMÉ. *Les terminologies d'entreprise se trouvent aujourd'hui rapprochées de représentations sémantiques issues de pratiques et de disciplines différentes, qui ont en commun la mise en relation des niveaux lexical et conceptuel au sein de structures informatiques. A partir du moment où ces représentations sont construites à partir de textes techniques reflétant les usages, leur rôle va bien au delà du simple inventaire terminologique ou définitoire. Désormais, elles sont les indices de convergences ou de dysfonctionnements (où), font ressortir des vocabulaires et des compétences métier (qui) ou encore des évolutions temporelles (quand). En cela, les représentations sémantiques construites à partir de documents sont des révélateurs utiles pour améliorer la communication et la gestion des connaissances dans l'entreprise. Pour mieux mesurer ces enjeux, il faut revenir sur l'importance des corpus sélectionnés, sur la nature des structures informatiques utilisées mais aussi sur les techniques et logiciels disponibles pour les construire à partir de textes.*

ABSTRACT. *Enterprise terminologies are now getting close to a panel of semantic representations coming from various disciplines and practices, and that share the relationship between lexical and conceptual levels within data structures. As long as these representations are built up from texts that reflect usage, their role goes far beyond drawing a simple inventory of terms and definitions. From now on, these structures are indicators of either convergences or misfunctions (where), of vocabularies as well as professional competences (who) or even temporal evolutions (when). By doing so, document based semantic representations are detectors useful to improve the company communication and knowledge management. To better measure these challenges, it is worth underlying the impact of corpora selection the exact nature of data structures and the kind of available techniques and software for their design from texts.*

MOTS-CLÉS : *terminologies, modélisation de connaissances, sémantique textuelle, ingénierie des connaissances, ontologies.*

KEYWORDS: *terminology, knowledge modelling, textual semantic, knowledge engineering, ontology.*

Actes de CIFT 2004, pages 33 à 39

 Actes de CIFT 2004, pages 33 à 39

1. Introduction

Les terminologies d'entreprise se trouvent aujourd'hui rapprochées de toute une gamme de représentations sémantiques issues de pratiques et de disciplines différentes, qui ont en commun la mise en relation des niveaux lexical et conceptuel au sein de structures informatiques. A partir du moment où ces représentations sont construites à partir de textes techniques reflétant les usages, leur rôle va bien au delà du simple inventaire terminologique ou définitoire. Désormais, elles sont les indices soit de convergences soit de dysfonctionnements par domaine spécialisé ou par équipe (où), de connaissances soit partagées soit implicites. Elles font ressortir des vocabulaires autant que des compétences métier (qui) ou encore des évolutions temporelles (quand). En cela, les représentations sémantiques construites à partir de documents sont des révélateurs, des supports pour améliorer la communication et la gestion des connaissances dans l'entreprise. Pour mieux mesurer ces enjeux, il faut revenir sur l'importance de la sélection des textes pour obtenir des corpus pertinents, sur la nature des structures informatiques actuellement utilisées pour rendre compte de terminologies sémantiques (terminologies, bases de connaissances terminologiques, bases de données lexicales et différents types d'ontologies) mais aussi sur les techniques et logiciels disponibles pour les construire à partir de textes (concordanciers, extracteurs de termes et de relations, basés sur des méthodes linguistiques ou sur des analyses statistiques). Nous ferons un rapide panorama de ces travaux (parties 1 et 2) avant de nous appuyer sur une série d'expériences menées dans différentes entreprises pour illustrer ces différentes dimensions (parties 3 et 4).

2. Une gamme de supports informatiques pour les terminologies

Historiquement, la diversité des ressources terminologiques s'explique par la volonté de répondre à des besoins propres à des types d'applications particulières. Ainsi, la plupart des bases de données terminologiques ont été définies pour faciliter le travail des traducteurs dans un domaine donné. Les langages documentaires sont les support de travail des documentalistes : ils leur fournissent un vocabulaire dans lequel seront puisés les mots-clés servant à indexer leurs collections. Autre exemple, les thesaurus d'entreprise servent avant tout de norme pour l'aide à la rédaction ou pour la communication.

Avec la mise sur support informatique de ces ressources terminologiques, l'arrivée de nouvelles applications mais aussi avec l'évolution des contenus des ressources, on assiste à une diversification des ressources disponibles : *thesaurus* pour indexation automatique (Roussey *et al.,* 2001) ; *réseaux de termes* et *construction d'index* pour moteurs de recherche spécialisés ; *index* hypertextuels pour documents électroniques ; b*ases de données terminologiques multilingues* pour

traduction automatique ; *bases de données lexicales* comme Wordnet comme aide à la traduction ou comme sources de méta-data pour annoter des documents ; *ontologies* pour systèmes à base de connaissances ou pour le Web Sémantique (Charlet *et al.,* 2003).

Cette diversité correspond à un renouvellement tout à fait intéressant. D'une part, par une sorte de convergence des besoins, des ressources prévues pour une utilisation sont détournées pour un nouvel usage. D'autre part, le caractère très mouvant des contextes d'usage rend très vite obsolètes des structures figées, fait ressortir des phénomènes linguistiques rendant plus complexes la mise au point de terminologies et souligne la nécessité de s'appuyer sur les usages réels des termes. Finalement, le support informatique n'est qu'une illusion d'uniformité, d'homogénéisation, qui ne doit pas faire oublier que ce sont bien les usages qui déterminent les modèles (Lainé-Cruzel, 2001).

Ces ressources ont en commun un modèle de données comportant à minima un réseau conceptuel plus ou moins formel (formé de concepts définissant des classes ou des instances, et des relations entre concepts définissant des attributs ou des propriétés) et /ou une réseau terminologique (termes décrits par des informations grammaticales ou sémantiques, reliés par des relations syntaxiques). Ces deux réseaux, lorsqu'ils cohabitent dans la même structure (comme dans WordNet ou dans les bases de connaissances terminologiques (Condamines, 2003)), sont reliés par des liens termes-concepts permettant de traduire de manière souple les différents sens de termes du domaine en fonction de leurs usages. Les ontologies sont la version actuelle la plus riche et souvent la plus formelle des réseaux sémantiques. A l'autre extrême, les lexiques ne comportent que des descriptions de termes.

Il nous semble important de bien définir la notion d'ontologie, car elle est utilisée aujourd'hui dans une multitude de contextes, souvent en lieu et place du nom d'autres ressources terminologiques. Or abuser de ce terme pour désigner des ressources qui n'ont rien d'ontologies conduit en ce moment à une confusion tout à fait négative. Cette confusion débouche par exemple sur l'utilisation de structure parfois trop riches ou trop complexes par rapport aux besoins réels, ou encore à exagérer la pertinence des ontologies alors que d'autres types de ressources sont tout aussi utiles. Finalement, la dissolution des particularités propres à chaque type de ressource fait perdre de vue leurs contextes d'usage, les connaissances qu'elles permettent de révéler ou encore les techniques à adopter pour les construire.

Les ontologies des modèles formels des connaissances d'un domaine pertinentes pour une application, une tâche donnée (Gomez-Perez *et al.,* 2003). Elles sont définies afin de disposer d'une représentation consensuelle, partageable et inter-opérable entre applications. Au-delà des bases lexicales ou des thesaurus, leur particularité est la rigueur et la précision de la conceptualisation des connaissances, répondant à l'application de règles de « bonne structuration » ou de normalisation. En particulier, la définition des concepts et des relations doit garantir leur unicité,

leur différenciation et l'unicité de leur interprétation formelle (sémantique) (Charlet *et al*., 2003).

3. Des terminologies aux textes et inversement

Les liens entre ressources terminologiques et textes ou même les documents contenants ces textes sont étroits tant à cause des usages prévus de ces ressources, souvent pour rédiger, consulter ou gérer des textes, qu'à cause de la richesse de la source de connaissances que les textes peuvent constituer à travers l'usage des termes. Ces liens sont renforcés dans le contexte des documents électroniques et du web, toujours selon les deux facettes de l'utilisation et de la construction de ces ressources : d'un côté, un nombre croissant d'applications documentaires fait désormais appel à des ressources terminologiques alors qu'en amont, un large éventail de logiciels permet de les construire plus rapidement à partir de textes, en conservant des liens entre textes et ressources (Condamines *et al*., 2000). De nombreux articles (Hamon *et al*., 2002) font le point des techniques statistiques et linguistiques à la base des outils de traitement automatique des langues utiles pour construire ces terminologies (extracteurs de termes ou de relations, concordanciers, identificateurs de classes sémantiques , …). Les travaux les plus récents en la matière portent justement sur le renforcement du lien entre textes et ressources, en utilisant l'apprentissage automatique pour retrouver en corpus de nouvelles instances de concepts, ou indexer des textes en repérant des formes lexicales révélant des concepts (Nazarenko *et al*., 2001) (Poibeau, 2001).

4. Les terminologies comme révélateurs d'identité et de savoir faire implicites

Les besoins des entreprises relatifs aux ressources terminologies correspondent en fait soit à la terminologie, soit à la gestion documentaire, soit aux connaissances. Concernant les documents, les entreprises se heurtent à de multiples difficultés comme l'existence encore importante du support papier, de gros volumes documentaires, de leur gestion dans le temps et de leur accès, de la capacité à en interpréter ou exploiter le contenu, etc. Les besoins exprimés sur les vocabulaires portent sur la définition de méthodes et d'outils pour contrôler les défaillances langagières, repérer des ambiguïtés ou des polysémies, etc. En matière de connaissances, les entreprises souhaitent mieux la véhiculer, la retrouver, la localiser, la rendre accessible et la ressource terminologique est tantôt perçue comme un médiateur renvoyant à la source que sont les documents, tantôt comme une nouvelle source de connaissances à exploiter directement (Dieng *et al.,* 2001). L'enjeux est la pérennisation des techniques, des savoir faire et des compétences.

Les réponses vont bien au delà de la structure de données produite. Tous les échanges et les interrogations nécessaires pour constituer une terminologie sont d'excellents révélateurs de la nature des connaissances détenues par les différents

acteurs de l'entreprise à différents moments de son existence (qui – où - quand) (Aussenac-Gilles *et al.,* 2003b). L'étude des cohérences ou divergences terminologiques, voire conceptuelles, conduit à en repérer les répercussions éventuelles, et d'anticiper les dysfonctionnements entre domaines spécialisés, métiers ou équipes. La mise en évidence de la part des connaissances partagées par rapport aux connaissances implicites constitue un bon indicateur. Ainsi, les analyses terminologiques contribuent à mieux caractériser l'identité de l'entreprise. Elles n'ont pas seulement un intérêt technique (modélisation de connaissances) ou organisationnel (meilleure coordination du travail, meilleure caractérisation des « cultures » des divisions ou encore support à la réalisation de tâches), elles ont également un intérêt pour le marketing (image interne et externe de l'entreprise).

4.1 Terminologies et identité des entreprises

Nous illustrons ces analyses à partir de trois études de cas, en insistant pour chacune sur un aspect particulier :

- Le vocabulaire comme médiateur d'un travail collaboratif : l'analyse terminologique met en évidence rapidement des termes ayant des sens différents d'une équipe à l'autre et conduisant à des incompréhensions. La nécessité de disposer d'un référentiel terminologique commun conduit à rendre compte des différents points de vue des équipes au sein de la terminologie partagée, de l'enrichir de définitions de groupes verbaux correspondant aux actions à côté des groupes nominaux, et enfin à l'utiliser pour former les équipes à une meilleur connaissance de leurs partenaires.
- le vocabulaire comme révélateur de l'image de l'entreprise : l'analyse des documents de communication interne et externe permet de restituer à leurs auteurs ce qu'ils font ressortir de l'activité et des compétences de l'entreprise, et de vérifier la cohérence avec l'image souhaitée.
- Rendre compte des composantes de l'entreprise via la terminologie : une étude différenciée des terminologies d'équipes métier ou de groupe liés à l'organisation de l'entreprise rend compte des écarts de vocabulaire mais aussi des sens différents donnés à des termes communs. De manière inattendue, l'analyse de textes produits par ces équipes restitue des rapprochements ne correspondant pas exactement au découpage organisationnel et révèle les véritables réseaux d'échange et de collaboration.

4.2. Les terminologies comme révélateurs de savoir-faire

Deux études de cas nous permettent d'illustrer la manière dont la mise au jour du vocabulaire de l'entreprise à travers les usages ainsi que la structuration conceptuelle associée révèlent ou donnent accès à des savoir-faire implicites.

- Des ontologies pour l'accès à des savoir-faire : alors que la mise en forme d'une ontologie est déjà un premier pas vers le repérage de connaissances implicites, son utilisation au sein d'un système de gestion des connaissances oblige la mise en correspondance entre différents types de données (textes, nomenclatures, ontologie, etc.) et rend ainsi explicite des savoir-faire nouveaux (Aussenac-Gilles et al., 2003a). Prévoir l'utilisation de la ressource au sein du système de gestion des connaissances requiert donc des analyses complémentaires et la mise en forme de connaissances supplémentaires.
- Une ressource terminologique modélisant des connaissances techniques jusqu'ici ni rassemblées ni explicitées : Dans le cadre de la construction d'un système de classification documentaire pour la veille technique et scientifique, la ressource élaborée pour guider la reformulation de requête fournit l'occasion à l'entreprise de mettre à plat ses compétences et savoir-faire dans un domaine jusque là peu structuré. La ressource est alors envisagée comme support à la mise en place d'une formation.

6. Conclusion

Nous avons montré qu'au delà des rôles désormais bien identifiés des ressources terminologiques comme vecteurs de connaissances, normes de vocabulaire, support à la communication, à la recherche documentaire et à la gestion de connaissances, ces ressources autant que le processus de leur construction à partir de textes jouent des rôles plus complexes de révélateurs de l'identité de l'entreprise et de ses composantes, de ses savoir-faire, de leur communication et de leur gestion. D'autres types d'études, plus sociologiques, associées à des projets terminologiques ou ontologiques devraient déboucher sur l'identification de nouveaux enjeux. Enfin, ces recherches font ressortir deux besoins en matière de ressources terminologiques et ontologiques. Tout d'abord, le paradoxe du caractère statique de ces ressources et du contexte très changeant de leur utilisation conduit à envisager des structures mises à jour ou reconstruites très régulièrement. Ensuite, une meilleure maîtrise des types de ressources nécessaires pour différents types d'applications ou d'objectif de caractérisation de l'entreprise (Bourigault *et al.,* 2004) devrait permettre d'éviter de considérer que la solution universelle sont les ontologies, et de redonner tout leur intérêt aux thesaurus, terminologies ou autres bases de connaissances terminologiques.**Bibliographie**

Aussenac-Gilles N., Biébow B., Szulman S. (2003a), Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In *Ingénierie des Connaissances*. R. Teulier, P. Tchounikine et J. Charlet Eds. Paris : L'Harmattan. A paraître en 2004.

Aussenac-Gilles N., Bourigault D., Teulier R. (2003b), Analyse comparative de corpus : cas de l'ingénierie des connaissances. Actes de IC2003 (14° journées Francophones d'Ingénierie des Connaissances). Présidente : R. Dieng-Kuntz. Laval (F), 1-3 Juillet 2003. Presses Universitaires de Grenoble. pp 67-84.

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. *Revue d'Intelligence Artificielle (RIA*). « Techniques Informatiques et Structuration de Terminologies ». Pierrel J.M. et Slodzian M. (Ed.). Paris : Hermès. Vol. 18. N°1/2004. pp 87-110.

Charlet J. Laublet P. Reynaud C., *Web sémantique*, rapport final de l'action spécifique 32 du CNRS/STIC. Déc. 2003.

Condamines A. (2003) : *Sémantique et corpus spécialisé : Constitution de bases de connaissances terminologiques.* Mémoire d'Habilitation à Diriger les Recherches, Juin 2003, Université Toulouse Le Mirail ; ERSS : *Carnets de grammaire*.

Condamines A., Rebeyrolle J, (2000) Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In J. Charlet, M. Zacklad, G. Kassel, D. Bourigault, (eds). : *Ingénierie des Connaissances, évolutions récentes et nouveaux défis.* Paris : Eyrolles. 2000, pp.225-242.

Dieng-Kuntz, R., Corby, O., Gandon, F., Giboin, A., Golebiowska, J., Matta, N. and Ribière, M. (2001). *Méthodes et outils pour la gestion des connaissances ; une approche pluridisciplinaire du Knowledge Management*. 2e édition, Dunod, Paris.

Gomez-Perez A., Manzano Macho D., (2003). *A survey of ontology learning methods and techniques*. Deliverable 1.5. IST Project IST-2000-29243 OntoWeb. May 2003. http://www.ontoweb.org/

Hamon T., Nazarenko A., (eds), (2002) : Structuration de terminologie. *TAL* volume 43 - n°1/2002.

Lainé-Cruzel S. (2001), Vers un nouveau positionnement des professionnels de l'information. *3ème colloque du Chapitre français de l'ISKO* (International Society for Knowledge Organisation) : Filtrage et résumé automatique de l'information sur les réseaux. Paris, 5-6 juillet 2001.

Nazarenko A., Zweigenbaum P., Habert B. et Bouaud J. (2001) Corpus-based Extension of a Terminological Semantic Lexicon, Recent Advances in Computational Terminology. John Benjamins, 2001.

Poibeau T. (2001), Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à nombre fini d'états, In *Actes de la Conférence Française de Traitement Automatique de la Langue*, (TALN'2001), 2001.

Roussey C., Calabretto, S., Pinon, J.-M. (2001). SyDoM: A multilingual Information Retrieval System for Digital Libraries. In *Proceedings of the 5th International ICCC/IFIP Conference on Electronic Publishing: ELPUB'2001*, Canterbury (UK), p. 150-160.