GLOTTOPOL

Revue de sociolinguistique en ligne

N° 8 – juillet 2006

Traitements automatisés des corpus spécialisés : contextes et sens

SOMMAIRE

PRESENTATION

Myriam Mortchev-Bouveret
Laboratoire CNRS Dyalang, Université de Rouen

Un numéro intitulé « traitements automatisés des corpus spécialisés : contextes et sens » peut surprendre les lecteurs de la revue Glottopol consacrée habituellement à la sociolinguistique. Il n'est en effet pas question ici de sociolinguistique mais de travaux menés sur corpus par des linguistes et informaticiens. L'Université de Rouen et le laboratoire DYALANG ont hébergé en 2005 la conférence TIA (cf TIA 2005) et ce numéro souhaitait poursuivre un peu les discussions amorcées de même que celles mises en route par une collaboration pluridisciplinaire menée lors de l'action CNRS ASTICCOT au sein des STIC (Aussenac-Gilles N. et Condamines A. 2003). Ce numéro traite de T.A.L et de terminologie, il est consacré aux traitements des langues spécialisées et présente des recherches menées sur corpus pour des visées telles que la traduction, l'acquisition lexicale, la recherche d'informations, la veille documentaire. Ce numéro ancré dans une université, l'Université de Rouen, où les travaux de sociolinguistique ont donné naissance à une approche socioterminologique (Gaudin 2003), a voulu présenter également un travail en cours recourant à une démarche socioterminologique en traduction informatisée multilingue (T.D. Tran). D'autres travaux rouennais sont en préparation dans cette perspective (cf. Baudouin *et al.* 2003). Cette démarche socioterminologique et informatique en est à ses débuts et intégrer la variation linguistique, les variétés de communautés selon une approche informatisée des corpus est une voie récente.

Les travaux présentés ici sont la rencontre de doubles, voire triples compétences et formations universitaires pour leurs auteurs : informatique, linguistique, domaines spécialisés, traduction.

Voilà donc l'esprit de ce numéro illustrant un champ de recherche largement pluridisciplinaire. Nous regrettons quelques articles perdus en cours de chemin, cités ici entre les lignes. Néanmoins, voici la livraison. Que les auteurs soient remerciés de leur collaboration.

Le numéro regroupe les articles suivants : Terminologie médicale bilingue anglais/français : usages clinique et législatif (Aurélie Névéol, Sylwia Ozdowska), Prendre en compte la dimension globale d'un corpus dans la contextualisation du sens : expérimentations en informatique linguistique (Pierre Beust, Thibault Roy), La conceptualisation métaphorique en biomédecine : indice de conceptualisation et réseaux lexicaux (Sylvie Vandaele, Sylvie Boudreau, Leslie Lubin, Elizabeth Marshman), Faire se rencontrer les parallèles : regards

croisés sur l'acquisition unilingue et multilingue (Pierre Zweigenbaum et Benoit Habert), Système de recherche d'information médicale par croisement de langues : vietnamien-français-anglais (Tran Duc Tuan). La traduction informatisée est un premier axe de travail. Elle est présente dans trois articles : le premier expose des recherches sur corpus parallèles en vue de traductions automatiques multilingues : langage biomédical et langage du droit pour traduction bilingue anglais-français (A. Névéol et S. Ozdowska), le second propose un travail en cours traitant de la recherche d'information médicale par croisement de langues français-anglais-vietnamien (T.D. Tran) ayant recours à la traduction alignée ; le troisième est un travail en construction, d'une chercheuse de Montréal et son équipe qui présentent ici une autre visée de la traduction informatisée : une base de données reposant sur le repérage de métaphores et la conceptualisatoin des indices pour une base de données biomédiale destinée aux traducteurs français-anglais.

Qu'il s'agisse de traduction ou d'une autre application, la constitution de corpus alignés, comparables, parallèles est une méthode à laquelle trois auteurs ont recours ici, A. Névéol et S. Ozdowska ainsi que T. D. Tuan ci-dessus. C'est aussi le procédé utilisé par P. Zweigenbaum et B.Habert dans un travail multilingue réalisé au sein de l'Inalco, concernant l'acquisition sémantique lexicale (semi-) automatique en contexte multilingue pour la constitution de dictionnaires sémantiques utilisant des corpus comparables.

Une autre dimension, celle de la variation, est présente dans trois articles. La variation prise en compte chez les communautés de locuteurs est explorée dans l'article de T. D. Tuan, c'est également la préoccupation de deux chercheurs informaticiens, spécialistes de T.A.L (P. Beust et T. Roy), qui développent une approche centrée autour des besoins d'un utilisateur ou d'un petit groupe d'utilisateurs. L'article se situe dans une visée différente qui n'est pas celle de la traduction comme c'est le cas dans l'article de T. D. Tuan, mais ils mettent en œuvre des traitements sémantiques adaptés à certaines tâches informatisées, interfaces de lecture rapide d'ensembles documentaires en particulier. L'article de Zweigenbaum et Habert quant à lui repose sur la notion de *types de ressources,* essentielle à la constitution de corpus comparables et nécessitant de situer les corpus selon leur genre textuel.

Le thème autour duquel est centré le numéro est nommé « Contextes et sens ». Quelles sont les difficultés posées par la constitution de ressources ou de modélisations linguistiques qui intègrent le contexte linguistique et extra-linguistique comme une dimension essentielle du fonctionnement linguistique des termes ? Comme le souligne Rastier dans un chapitre intitulé « La lexie en contexte : de la signification au sens » (Rastier, 1994 : 68) :

> *« En passant de la lexie comme contexte à la lexie en contexte, nous ne quittons pas la syntagmatique. On retrouve entre les mots les mêmes types de relations contextuelles que l'on discerne entre les morphèmes, ce qui montre tout à la fois combien est arbitraire la frontière du mot et combien utile une typologie des relations contextuelles. Il est en outre douteux que le mot soit perçu isolément autant pour son contenu que pour son expression. Nous formulons l'hypothèse qu'il en va de même, corrélativement, pour le signifié des mots, qui serait perçu par des activations contextuelles. »*

Cette position théorique a donné naissance au courant de la *terminologie textuelle* (Slodzian, 2000) se penchant précisément sur une typologie des relations contextuelles en vue du traitement informatisé des données terminologiques. Comment donc définir le contexte

concerné dans les articles présentés ici ? La définition suivante s'y applique en partie mais ne suffit pas :

> *« Par rapport à un élément quelconque d'une suite linguistique, le contexte est l'ensemble des unités qui le précèdent et le suivent. Le contexte pris en considération reçoit des limitations proportionnelles au statut et à la dimension de l'unité concernée : le contexte d'un phonème sera la syllabe (éventuellement le morphème), le contexte du morphème, le syntagme, celui du syntagme, la phrase. Pour la phrase, le contexte est constitué par des unités discursives dont la délimitation s'opère selon des procédures qui ne relèvent plus exclusivement de la linguistique » (Arrivé M., Gadet F. et Galmiche M., 1986 :185).*

Cette autre définition la complète :

> *« Le contexte est l'ensemble des éléments situationnels extra-linguistiques au sein desquels se situe l'acte d'énonciation de la séquence linguistique. En ce second sens, contexte renvoie à référent. » (ib.).*

L'extra-linguistique dans les articles recueillis ici concerne la variation mais aussi le contexte de production. Comment prendre en considération les communautés de locuteurs, « la situation de production et d'interprétation » (Condamines, 2005 : 33) ? « *Un corpus étant constitué de textes ou d'extraits de textes, il est difficile de faire totalement l'impasse sur le fait que ces textes ont été rédigés dans des situations particulières qui impliquaient des protagonistes ayant des intentions particulières* » (*ib.*). Le contexte peut donc aussi s'envisager comme « construction et interprétation du sens par des sujets », « intertexte » : *« La question du sens (sa construction et sa nature) est bien sûr très liée aux rapports entre des documents (majoritairement textuels) et des sujets interprétants »* (Beust et Roy : ci-inclus).

D'autres éléments interviennent dans une perspective de construction du sens en contexte concernant le présent propos, « les traitements automatisés de corpus spécialisés » :

- Quels sont les éléments syntaxiques de construction du sens à considérer dans le contexte ? Règles de sous-catégorisation, marqueurs (prépositions, affixes, suffixes, préfixes, syntagmes, etc.), contraintes de sélection ? Mais comme le soulignent Zweigenbaum et Habert (ci-inclus), «*ne pas se cantonner aux traits syntaxiques » : « (…) deux extrémités possibles pour la représentation des contextes d'un mot. La première, «pauvre », se contente de repérer de simples cooccurrences entre mots, dans une fenêtre textuelle considérée comme un « sac de mots », c'est-à-dire en perdant l'ordre des mots entre eux. La seconde bénéficie d'une analyse syntaxique, même partielle, et repose sur les dépendances syntaxiques élémentaires entre mots* ».

- Quelles sont les affinités sémantiques et syntaxiques entre les unités ? Le sens d'une unité linguistique est constituée de ses relations contextuelles également définies ainsi par Cruse dans un chapitre intitulé *A contextual approach* : *« We can figure the meaning of a word as a pattern of **affinities** and **disaffinities** with all the other words in the language with which it is capable of contrasting semantic relations in grammatical contexts. Affinities are of two kinds, **syntagmatic** and **paradigmatic** »* (Cruse, 1986 *:* 18).

- Lors de l'interprétation de l'énoncé le sens est-il global ou compositionnel ? Comment doit-on ainsi interpréter, consigner, modéliser la phraséologie, les collocations ? C'est l'objet en particulier des travaux de l'équipe OLST (cf. Orliac B. 2006).

- Comment considérer des éléments cognitifs de l'interprétation telles les métaphores et comment les modéliser ? (cf. Beust P. et Roy T., ci-inclus, cf. Vandaele S. ci-inclus)

On se doit donc dans une perspective sémantique d'élargir la notion de contexte linguistique et extra-linguistique à celle de contexte d'interprétation, voire de « calcul du sens et perception sémantique » comme le montre l'article de Zweigenbaum et Habert (ci-inclus).

En conclusion, si les questions concernant la nature des termes et des concepts terminologiques étaient au cœur de la réflexion de la décennie 1990-2000, envisageant le sens du point de vue de sa représentation ; les questions liées au sens, aux contextes et aux corpus émergent dès 2000 (cf. Béjoint et Thoiron (dir.) 2000, Bourigaut, Jacquemin et L'Homme 2001, AUF 2005, Condamines 2005) et soulèvent alors les problèmes liés à son interprétation, à sa modélisation. Dans cette perspective, les corpus et les contextes ne peuvent pas être envisagés comme de simples preuves langagières, mais comme un élément de la construction du sens et constituent en cela un défi à la question du sens en langue. Selon le programme dessiné par la terminologie textuelle, c'est donc bien à une typologie des relations contextuelles que les terminologues-informaticiens doivent s'attacher afin d'approfondir la question de la modélisation du sens dans les langues spécialisées.

Bibliographie

Arrivé M., Gadet F. et Galmiche M., 1986, *La grammaire d'aujourd'hui*, Flammarion

AUF, 2005, Agence Universitaire de la francophonie, *Mots termes et contextes*, 7èmes journées scientifiques, Réseau de chercheurs Lexicologie, terminologie et traduction, ISTI, Bruxelles, 8-10 septembre 2005

Aussenac-Gilles N. et Condamines, A., 2003. *Rapport final de l'action spécifique « Corpus et Terminologie »*, http://www.irit.fr/ASSTICCOT

Baudouin N., Holzem M., Saidali Y., Labiche J., 2003, « Modélisation des connaissances et construction d'un consensus : apport de la socioterminologie à une plate-forme en traitement d'image », dans *Actes de la conférence TIA 2003*, p. 54-68

Béjoint H. et P. Thoiron (dir.), 2000, *Le sens en terminologie*, Presses Universitaires de Lyon

Bourigault D., Jacquemin C. et L'Homme M.-C., 2001, *Recent advances in computational Terminology*, John Benjamins Publishing Company, Amsterdam-Philadelphia

Condamines A. (dir.), 2005, *Sémantique et corpus*, Hermès-Lavoisier

Cruse D.A., 1986, *Lexical Semantics*, Cambridge University Press

Gaudin F., 2003, *Socioterminologie. Une approche sociolinguistique de la terminologie*, coll. « Champs linguistiques », éd. Duculot, Louvain-la-Neuve

Habert B., Nazarenko A. et Salem A., 1997, *Les linguistiques de corpus*, Armand Colin

Slodzian M., 2000, « L'émergence d'une terminologie textuelle et le retour du sens », dans Béjoint H.et Thoiron P.(dir.), 2000, p. 61-85

Orliac B., 2006, « Colex : un outil d'extraction de collocations spécialisées basé sur les fonctions lexicales », dans *Processing of Terms in Specialized Dictionaries*, L'Homme M.-C. (ed.), p. 261-280

Rastier F., Cavazza M. et Abeillé A., 1994, Sémantique pour l'analyse. De la linguistique à l'informatique, Masson

TIA, 2005, *Actes de la conférence TIA 2005, Terminologie et Intelligence artificielle*, DYALANG-Université de Rouen, Mont Saint Aignan 4 et 5 avril 2005, http://tia.loria.fr/

GLOTTOPOL

Revue de sociolinguistique en ligne