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/

TERMINOLOGIE MÉDICALE BILINGUE ANGLAIS/FRANÇAIS : USAGES CLINIQUE ET LÉGISLATIF

Aurélie Névéol,

Équipe CISMeF & CGSIS, CHU de Rouen
National Library of Medicine, Bethesda

Sylwia Ozdowska

ERSS - CNRS & Université de Toulouse le Mirail

1. Introduction

La santé est l'un des domaines de spécialité les plus dotés en terminologies destinées à des usages aussi variés que le codage des dossiers patients ou la description de documents d'information en santé. Beaucoup de ces terminologies sont développées en anglais et les ressources disponibles dans d'autres langues comme le français demandent à être complétées. Ainsi, plusieurs projets récents tels que UMLF (Zweigenbaum *et al.*, 2003) et VUMeF (Darmoni *et al.*, 2003) ont abordé le développement ou l'enrichissement de terminologies médicales en français par des approches basées sur le traitement automatique de corpus du domaine.

Pour les terminologies issues d'une langue autre que le français, la traduction des termes relève d'une double compétence, à la fois en traduction – afin de rester fidèle à la terminologie originelle – et dans le domaine de spécialité – afin de s'assurer que le terme issu de la traduction correspond bien au concept à désigner. La disponibilité de linguistes ou de terminologues maîtrisant ces deux aspects étant réduite, il apparaît opportun de séparer les deux composantes du problème, et de traiter successivement la traduction des unités terminologiques puis la validation des traductions obtenues avant inclusion dans la terminologie.

Dans ce cadre, nous avions proposé et mis en œuvre une méthode de traduction automatique de termes du domaine médical à l'aide de corpus parallèles (Névéol et Ozdowska, 2005 ; Ozdowska *et al*., 2005). Ainsi, nous avions défini deux contraintes pour nos corpus de travail : d'une part la *qualité de la traduction* des textes parallèles, et d'autre part la *couverture des termes* que nous cherchions à traduire. Pour ce dernier point, nous nous étions naturellement orientées vers des corpus attestés du domaine médical, constitués de Résumés des Caractéristiques du

Produit (corpus RCP) ou de textes issus des sites officiels de diverses institutions de santé canadiennes (corpus CISMeF[1]). Quelques essais effectués sur des extraits d'un corpus non spécialisé mais traitant néanmoins de sujets liés à la santé, le Hansard, nous avaient permis de constater que ce type de corpus pouvait également satisfaire à nos critères, tant au niveau de la qualité de la traduction que de la couverture des termes à traduire. Nous avions pu observer à cette occasion des divergences dans les traductions extraites du Hansard par rapport à nos corpus médicaux. Pour certains termes polysémiques, l'un des sens était plus représenté dans le Hansard que dans les corpus médicaux, voire exclusivement représenté dans le Hansard. Par exemple, pour le terme *drug*[2], la traduction «drogue » était plus fréquente dans le Hansard, alors que c'est la traduction « médicament » qui était majoritaire dans le corpus RCP.

Ces différences nous ont amenées à nous poser la question de la projection d'une terminologie spécialisée renvoyée par chacun des corpus : les traductions extraites pour un même terme sont-elles complémentaires ? Les usages renvoient-ils à différents concepts dénotés par les mêmes termes ? L'ensemble des usages qui se dégagent des différents contextes doivent-ils être pris en compte dans la terminologie ?

Dans cet article, nous nous proposons d'étudier ces questions dans le contexte de notre problématique de traduction des synonymes de la terminologie MeSH® (Medical Subject headings) à partir d'un corpus spécialisé du domaine médical et d'un corpus non spécialisé à forte coloration juridique. Notre objectif est double : il s'agit d'une part de dégager les apports de ces deux contextes à la terminologie MeSH, et d'autre part d'apprécier dans quelle mesure une telle étude permet de tirer des conclusions sur les critères de choix d'un corpus (parallèle) pour l'enrichissement d'une terminologie (bilingue). Après une description de la terminologie médicale MeSH ainsi que des deux corpus de travail, CESART et le Hansard, nous examinerons la couverture des termes à traduire d'un point de vue quantitatif et qualitatif. Puis, nous nous intéresserons à la traduction des termes couverts à travers une étude de cas. Enfin, nous discuterons les résultats obtenus afin de dégager des éléments de réponse aux questions soulevées par cette étude.

2. Le MeSH (Medical Subject Headings)

Le thésaurus MeSH est la terminologie de référence utilisée pour la recherche d'information dans le domaine bio-médical. L'ensemble des documents recensés dans la base documentaire MEDLINE sont indexés à l'aide de descripteurs MeSH. Créé dans les années soixante par la National Library of Medecine, ce thésaurus est mis à jour chaque année. La version 2005 comporte environ 23 000 mots clés, ainsi que 83 qualificatifs qui peuvent leur être associés afin de préciser le sens des concepts dénotés. Le thésaurus original (c'est-à-dire en anglais) comprend également près de 30 000 synonymes dont 25 000 ne sont pas traduits en français. Les mots clés MeSH (et les synonymes qui leur sont associés) renvoient à différents concepts du domaine médical, répartis en seize arborescences.

[1] Les documents bilingues rassemblés dans ce corpus étaient référencés dans le Catalogue et Index des Sites Médicaux Francophones (CISMeF) accessible sur http://www.cismef.org

[2] Tout au long de cet article, nous avons adopté les notations suivantes pour les différents termes dont nous discutons. Les termes MeSH sont représentés entre chevrons et en italique : <*terme MeSH*>, les synonymes MeSH en anglais – c'est-à-dire les termes que nous cherchons à traduire, sont représentés en italique : synonyme EN et les traductions extraites de nos corpus sont représentées entre guillemets : « synonyme FR traduit ».

Les exemples du Tableau 1 illustrent la variété de ces termes. Ils recouvrent en premier lieu les <*substances chimiques et médicamenteuses*> (arborescence D, 34% des mots clés), les <*maladies*> (arborescence C, 23% des mots clés) et les <*organismes*> (arborescence B, 11% des mots clés) mais aussi les termes liés à l'<*anatomie*> (arborescence A, 5% des mots clés), les <*soins de santé*> (arborescence N, 4% des mots clés) ou les <*personnes*> (arborescence M, 0,5% des mots clés).

Arbo. MeSH	Mot clé MeSH anglais	Mot clé MeSH français	Synonyme MeSH anglais (à traduire)
D	cardiovascular agents	agents cardiovasculaires	cardiovascular drugs
G	cell division	division cellulaire	cytokineses
A, J	milk, human	lait femme	breast milk
C	skin diseases, vesiculobullous	dermatoses bulleuses	sneddon wilkinson disease
N, J	ambulance	ambulance	mobile emergency unit

Tableau 1 – Extrait du MeSH

3. Choix et présentation des corpus de travail

3.1. Corpus spécialisé

Habert *et al*. (1997 : 38) définissent un corpus spécialisé comme étant un corpus «restreint à une situation de communication, un domaine, une langue de spécialité[3], c'est-à-dire un langage spécifique, très contraint du point de vue lexical, syntaxique, voire textuel, que l'on trouve dans les domaines scientifiques et techniques. »

Pour cette étude, nous avons considéré les thématiques abordées et les contextes de discussion de ces thématiques comme critères principaux de sélection des corpus parallèles : nous souhaitions étudier des thématiques relevant de la biomédecine telles qu'elles sont abordées dans deux contextes distincts. Le premier contexte concerne une situation où des experts du domaine s'adressent à des non experts. Le second concerne une situation où des non experts, éventuellement spécialisés dans un autre domaine, s'adressent à des non experts.

3.2. CESART

Ce corpus a été constitué à partir du site bilingue anglais/français Santé Canada (http://www.hc-sc.gc.ca) dans le cadre de la campagne d'évaluation CESART[4]. Il rassemble des publications du ministère de la santé du Canada à l'intention des citoyens canadiens, en particulier afin de les informer sur la prévention de diverses maladies et de les encourager à adopter un mode de vie sain.

Dans le cadre de notre étude, nous n'avons utilisé qu'un sous-ensemble du corpus initial. Nous n'avons, dans un premier temps, conservé que les documents pour lesquels la partie anglaise et française ne présentait pas un décalage de plus de deux paragraphes. Un calcul du nombre de paragraphes par document dans chaque partie avait révélé des différences importantes de ce point de vue. Certains paragraphes du document anglais, respectivement français, n'avaient pas de contrepartie dans l'autre langue, ce qui pouvait perturber l'alignement automatique au niveau des

[3] Pour une discussion sur la notion de « langue spécialisée », nous renvoyons le lecteur à Hamon (2000 : 25)
[4] Une présentation détaillée de cette campagne est faite sur http://www.technolangue.net/article58.html.

phrases nécessaire à la recherche d'équivalents. Les documents retenus ont fait l'objet d'un découpage et d'un alignement automatiques au niveau des phrases à l'aide de l'outil Japa (http:// http://rali.iro.umontreal.ca/Japa). Ensuite, une fois ce dernier effectué, nous n'avons retenu que les couples de phrases telles que la partie anglaise contenait au moins un synonyme MeSH à traduire. Au final, le bitexte[5] compte au total environ 1 million de mots pour 16 318 phrases alignées.

3.3. Le Hansard

Le Hansard est le Journal des débats à la Chambre des communes du Parlement canadien en anglais et en français[6]. Ces débats quotidiens sont enregistrés, transcrits et traduits au sein du service parlementaire du Bureau de la traduction, une agence du gouvernement du Canada, garant de la qualité de la traduction. Bien que les sujets abordés soient très variés, politique locale et mondiale, économie, société, santé, etc., ils sont traités du point de vue de la législation canadienne en vigueur. Le corpus Hansard rassemble ainsi plusieurs années de débats qui correspondent à plusieurs dizaines de millions de mots dans chaque langue. Le contexte dans lequel ce corpus est produit lui confère une nette coloration juridique. Cependant, la diversité des thématiques abordées fait qu'il est habituellement considéré comme un corpus général. Comme pour CESART, une partie alignée au niveau des phrases a été prélevée sur l'ensemble de ce corpus. En effet, comme pour le corpus CESART, nous avons sélectionné les segments alignés tels qu'au moins un synonyme MeSH dont on cherche la traduction était présent dans le segment. Les passages sélectionnés relèvent donc souvent de la santé publique, mais également d'autres thématiques (économie, …) comme l'illustrent les exemples cités au cours de l'article. Le bitexte ainsi obtenu contient près de 3 millions de mots pour 24 653 phrases alignées.

3.4. Caractéristiques générales

Les documents publiés par Santé Canada sont rédigés dans l'une des langues officielles du Canada (français et anglais). Il en est de même pour les débats rassemblés dans le Hansard qui ont lieu en anglais et en français. Ainsi, chacune de ces deux langues peut être tour à tour langue source ou langue cible. Cette information n'est pas disponible dans les versions du corpus dont nous disposons, ce qui rend la distinction impossible. C'est donc par abus de langage que nous parlerons de traductions françaises des synonymes anglais, les rôles pouvant être inversés de sorte que ce que nous appelons traduction peut en réalité correspondre à la version originale du terme.

Par ailleurs, si l'on s'en tient à une caractérisation du niveau de spécialisation des corpus en termes d'émetteur et de récepteur, ces deux corpus présentent un niveau de spécialisation que nous qualifierons de faible, mais à des degrés différents. En effet, dans les deux cas le récepteur, autrement dit le public visé, n'est pas un spécialiste du domaine. Par contre, le ou les émetteur(s), autrement dit le ou les auteur(s), sont des spécialistes du domaine de la médecine dans le cas de CESART et des non spécialistes pour ce qui est du Hansard.

Enfin, il s'agit de deux corpus canadiens, ce qui permet de gommer les biais qui auraient pu exister si des variantes régionales différentes des langues de travail avaient été comparées.

Le Tableau 2 reprend les caractéristiques générales des deux corpus d'étude.

[5] Corpus parallèle dont les segments en relation de traduction (ici les phrases) ont été mis en correspondance.
[6] http://parl.gc.ca/common/Chamber_House_Debates.asp?/Language=F

	CESART		Hansard	
Nombre de mots	**EN**	**FR**	**EN**	**FR**
	446 433	537 780	1 567 741	1 700 568
	984 213		3 268 309	
Nombre de phrases alignées	24 653		16 318	

Tableau 2 - Quelques caractéristiques générales des corpus CESART et Hansard

4. Couverture des termes à traduire dans les corpus

4.1. Aspect quantitatif

Au total, dans le MeSH 2005, 25 111 synonymes anglais ne sont pas traduits en français. Nous avons sélectionné un échantillon composé des 2 500 premiers de cette liste classée par ordre alphabétique des mots clés MeSH anglais correspondant aux synonymes à traduire. Dans un premier temps, nous avons limité notre travail à cet échantillon en raison des ajustements manuels nécessaires aux traitements détaillés en section 5. Sur les 2 500 synonymes de notre échantillon, 208 sont présents dans les deux corpus. Par ailleurs, 127 synonymes supplémentaires sont présents exclusivement dans CESART, et 61 autres synonymes sont présents exclusivement dans le Hansard. Globalement, la couverture est d'environ 70% – ce qui est supérieur à la couverture obtenue précédemment grâce aux corpus RCP et CISMeF.

Comme le montre le Tableau 3, il y a moins de synonymes différents dans le Hansard que dans CESART, par contre la fréquence (*i.e.* le nombre d'occurrences) des synonymes est plus importante dans le premier que dans le second. Ainsi, les synonymes de fréquence 1 représentent 30% des synonymes présents dans CESART et 20% dans le Hansard. Dans l'ensemble, la majeure partie des synonymes à traduire a une fréquence basse (inférieure ou égale à 10) dans les deux corpus, soit 72% dans CESART et 61% dans le Hansard. Par ailleurs, le nombre de mots dans le Hansard étant pratiquement trois fois supérieur à celui du corpus CESART, la densité en terminologie médicale, en termes de synonymes distincts et non de fréquence, est beaucoup plus élevée dans le second cas que dans le premier, ce qui va dans le sens d'un degré de spécialisation plus important pour CESART.

	CESART	Hansard
Nombre de synonymes présents	335	269
Fréquence moyenne des termes à traduire	46	125
Fréquence = 1	101 (30%)	55 (20%)
2 £ Fréquence £ 10	141 (42%)	101 (41%)
Fréquence > 10	93 (28%)	113 (39%)
Fréquence maximale	1813	4302

Tableau 3 - Nombre et fréquence des synonymes MeSH par corpus

De plus, si on examine les 20 synonymes les plus fréquents dans chacun des deux corpus (Tableau 4), on constate, d'une part, que seuls 4 synonymes, qui ne relèvent pas spécifiquement du domaine de la médecine, sont communs aux deux listes : *cost*, *organization*, *children*, *parent*. D'autre part, les synonymes les plus fréquents confirment l'orientation thématique globale de

chacun des corpus : celle de la santé pour ce qui est de CESART avec des termes tels que *virus*, *physician*, *laboratory*, *pharmacist*, *injection*, *tissue*, *cell*, etc. ; celle de la législation (et, par extension, des groupes d'individus concernés) pour le Hansard avec des termes tels que *pension*, *accident*, *aid*, *rights*, *human rights*, *prison*, *treaty*, *veteran*, *immigrants*, etc.

CESART		**Hansard**	
children	1813	cost	4302
virus	1441	children	3588
physician	1106	rights	3293
public health	1025	organization	1798
organization	1009	pension	1589
cost	822	industries	1031
foods	511	parent	883
laboratory	412	immigration	863
health service	379	hearings	745
parent	375	human rights	631
health promotion	355	aid	626
cell	289	immigrant	601
pharmacist	280	newspaper	537
mental health	248	accident	487
risk factor	236	suggestions	469
accountability	227	veteran	416
injection	143	treaty	375
tissue	132	prison	375
health status	131	waters	371
infants	114	research and development	322

Tableau 4 - Les 20 synonymes MeSH les plus fréquents par corpus

4.2. Aspect qualitatif

Le Tableau 5 ci-dessous présente la proportion des catégories de termes MeSH dans la terminologie, dans l'échantillon de synonymes à traduire (2ème colonne), et dans les synonymes présents dans les deux corpus (3ème colonne).

On constate que dans notre échantillon de synonymes à traduire, les catégories B et D sont sous-représentées, et les catégories G, H, I, J, M et N sont sur-représentées. On observe des écarts similaires de représentation dans les corpus pour ces catégories.

En revanche, pour les catégories C et N représentées de manière relativement similaire dans le MeSH et dans l'échantillon de synonymes, on observe une sous-représentation significative dans les corpus pour la catégorie C (<*maladies*>), et une sur-représentation également significative pour la catégorie N (<*soins de santé*>). Ces différences semblent conforter l'hypothèse du faible degré de spécialisation des corpus.

Arbo MeSH	MeSH (%)	Synonymes (0-2 500) (%)	Corpus (H+C) (%)
A	5,3	4,1	4,4
B	10,9	2,8	1,3
C	22,6	25,6	5,3
D	34,3	9,5	6,8
E	7,9	12,9	8,6
F	2,3	3,3	3
G	6,5	15	16,6
H	1,4	3,7	4,2
I	1,2	5,7	13,4
J	0,8	2,8	3,7
K	0,5	1,2	2,2
L	1	1,7	3,5
M	0,5	2,1	4,7
N	3,5	9,6	22,3

Tableau 5 - Couverture des catégories de termes MeSH par les corpus

Si on considère les termes présents dans les corpus pris séparément, il y a peu de différence dans la répartition des différentes catégories de termes. CESART comporte une proportion légèrement plus élevée de termes spécialisés (B- <*organismes*>, C- <*maladies*> et D- <*substances*>) et le Hansard une proportion légèrement plus élevée de termes généraux (N, <*soins de santé*> et I, <*phénomènes sociaux*>) au détriment des termes spécialisés (B, C et D). Ainsi, on peut dire que les termes présents dans le Hansard mais pas dans CESART sont :

- des termes appartenant à un domaine de la médecine sur lequel le droit statue (*abortion technics*, *anti-abortion group*) ;
- des termes relevant plutôt de l'économie (*healthcare economics and organizations*, *multiregional analyses*) ou des personnes (*workmen compensation*, *navy personnel*, *patients' visitor*) ;
- des termes généraux (*research technics*, *nautical medicine*).

Les termes présents dans CESART mais pas dans le Hansard sont au contraire plus techniques : *micronutrient*, *intestinal mucosa*, *ventricular tachycardia*, etc. Pour ce qui est des termes communs, on observe une fréquence beaucoup plus élevée des termes juridiques tels que *criminal justice* ou *human rights* dans le Hansard. Inversement, la fréquence des termes techniques tels que *medical records*, *e coli* ou *nervous system* est supérieure dans CESART.

5. Repérage des traductions

Pour un synonyme MeSH donné, le repérage de(s) traduction(s) consiste à identifier dans un couple de phrases alignées du corpus le ou les mots qui forment son équivalent français[7]. Ce repérage est effectué de manière semi automatique. Nous utilisons tout d'abord la méthode d'appariement syntaxique décrite dans (Ozdowska, 2004). Elle permet de repérer les traductions qui répondent à des schémas syntaxiques de correspondance standard entre l'anglais et le français tels que ceux décrits entre autres dans (Daille, 1994 ; Gaussier, 2001) et exploités spécifiquement

[7] Le repérage du terme de départ (synonyme MeSH anglais) est également réalisé de manière automatique.

en extraction de terminologies bilingues anglais/français. Ces schémas standard mettent en correspondance les structures comme Adj N et N Adj de l'anglais et du français, N N et N Adj ou encore N N et N prep N. Ainsi, pour le terme *nuclear energy*, nous sommes en mesure de repérer, dans les couples de phrases alignées, la traduction « énergie nucléaire ». Il en va de même pour les traductions standard des termes tels que *smoking cessation*, *family violence*, *health care reforms* qui sont «renoncement au tabac », « renoncement au tabagisme », « abandon du tabac », « violence familiale », « violence dans la famille », «réformes des soins de santé ». Par contre, la méthode d'appariement utilisée ne permet pas pour le moment d'identifier des traductions en dehors de ces schémas syntaxiques. Par exemple, le synonyme *smoking cessation* peut être traduit par « renoncer au tabac » ou encore «cesser de fumer » ; *family violence* admet comme autres équivalents « violence en milieu familial », « violence faite aux familles » ou encore « familles où sévit la violence » ; enfin *nuclear energy* peut être rendu par le nom générique « nucléaire » qui constitue un cas de réduction terminologique en contexte. Le repérage de toutes les traductions possibles ne se justifie pas nécessairement dans un objectif d'extraction de terminologie bilingue comme en témoignent les travaux de référence dans le domaine, notamment ceux cités ci-dessus, qui se concentrent sur les types de correspondance les plus fréquents. Dans la mesure où nous nous intéressons à l'usage des termes médicaux dans des communautés présentant différents niveaux de spécialisation dans le domaine ainsi qu'au lien entre usage et intégration dans une terminologie du domaine, il était important de disposer aussi bien des traductions standard que non standard. C'est pourquoi nous avons complété l'étape de repérage automatique par une phase de correction et de repérage manuel dans les phrases où la traduction correcte n'avait pu être trouvée automatiquement. À l'issue de ces deux phases de repérage, c'est ainsi en moyenne 2,2 équivalents par synonyme qui ont été repérés dans CESART et 3,5 dans le Hansard (Tableau 6). Ce dernier présente par conséquent un taux de variation dans les traductions plus important.

Dans un domaine de spécialité, le statut de *terme* confère théoriquement à une expression un statut monosémique, le sens du terme étant celui du concept qui lui est associé dans la terminologie considérée. L'existence de multiples traductions pour les synonymes que nous cherchons à traduire peut révéler autant de manières d'exprimer un même concept. Ainsi, un nombre plus élevé de traductions dans le Hansard peut être révélateur d'une plus grande variabilité terminologique dans ce corpus. À l'opposé, cela peut également résulter d'un plus faible degré de spécialisation, caractérisé par une résurgence de la polysémie des expressions. Nous présentons dans les sections 6 et 7 des études de cas de variation.

	CESART	**Hansard**
Nombre moyen de traductions extraites par terme	2,2	3,5

Tableau 6 - Nombre moyen d'équivalents par synonyme

6. Termes communs aux deux corpus

Nous proposons ici une étude approfondie des traductions extraites des deux corpus pour quelques termes ayant une fréquence similaire dans CESART et dans le Hansard. Les termes choisis sont significatifs des différences observées dans les deux corpus.

6.1. drug cost

Deux traductions du terme *drug cost* reviennent fréquemment dans les deux corpus; il s'agit de « prix des médicament » et de « coût des médicament ». La première, «prix des médicaments », correspond à des contextes où les entreprises pharmaceutiques ou les hôpitaux doivent prendre une décision concernant la facturation des médicaments aux patients (C1). A l'inverse, la seconde traduction, « coût des médicaments », est utilisée dans les contextes prenant en compte la charge financière liée aux médicaments pour le patient ou l'hôpital (C2). Dans le Hansard, le rôle du gouvernement dans les deux contextes – détermination des prix, prise en charge des coûts – est central. Ainsi, la distinction entre les utilisations de « prix » et « coût » est plus floue (H1, H2). Cette vision plus globale de la question donne lieu dans certains cas à l'emploi de désignations globales telles que « frais pharmaceutiques » (H3).

Cet exemple montre en outre que la manière globale dont le thème est traité dans le Hansard est dans ce cas la source d'une plus grande diversité lexicale. Dans CESART, deux contextes précis correspondant au cadre clinique « drug cost » sont représentés, et seuls les deux termes dénotant chacun de ces contextes peuvent donc apparaître dans le corpus. Pour ce synonyme, l'ensemble des traductions rencontrées dans les deux corpus peuvent être validées et intégrées à la terminologie.

C1 Le prix du médicament soumis correspondra au nombre de milligrammes administrés et non au volume administré.

C2 Le rapport du Comité reflète les préoccupations des Canadiens et des Canadiennes au sujet du coût des médicaments et de l'incidence de celui-ci sur le système de soins de santé.

H1 En outre, le gouvernement consulte les provinces pour s'assurer de garder les coûts des médicaments à un niveau raisonnable.

H2 Elles ont trait au projet du gouvernement de majorer les prix des médicaments en présentant le projet de loi sur lequel nous nous prononcerons aujourd'hui.

H3 Grâce à elle, ces derniers avaient notamment économisé en 1983 seulement quelque 211 millions de dollars en frais pharmaceutiques.

	CESART (20)	**Hansard (26)**
drug cost(s)	*coût(s) des médicaments coût du médicament coûts liés aux médicaments prix du médicament*	*prix des médicaments coût(s) des médicaments frais pharmaceutiques*

Tableau 7 – Traductions du synonyme drug cost(s)

6.2. birth weight/low birth weight

Le synonyme *birth weight* apparaît toujours accompagné d'un modifieur dans les deux corpus : dans la plupart des cas il s'agit de *low* pour les deux corpus mais on trouve également *high* et *unhealthy* dans CESART ainsi que *unsatisfactory* dans le Hansard[8]. Le Tableau 8 reprend les différents composés formés avec *birth weight* et leurs traductions dans chacun des corpus. On remarque que la présence d'un modifieur peut rendre l'extraction de la traduction du synonyme délicate dans la mesure où c'est la séquence modifieur + *birth weight* qui constitue alors l'unité

[8] Remarquons la différence de registre des modifieurs: "high" et "unhealthy" sont des termes cliniques neutres, alors que "unsatisfactory" est un terme plus général faisant appel à un jugement de valeur.

de traduction. Ainsi, dans le Hansard, *low birth weight* est rendu par « insuffisance pondérale », « insuffisance pondérale à la naissance » ou encore « petit poids » dans le contexte « bébés de petit poids », tandis que *unsatisfactory birth weight* dans le contexte *have an unsatisfactory birth weight* correspond à « n'ont pas un poids satisfaisant à la naissance ». Il n'y a guère que le dernier contexte qui est susceptible de fournir une traduction de *birth weight* exploitable dans une optique d'intégration à une terminologie, à savoir « poids à la naissance ». On constate qu'il existe une meilleure régularité dans les traductions de CESART. Quel que soit le modifieur, il est possible dans la majeure partie des cas d'isoler la traduction de *birth weight*. Ainsi, *low birth weight* est rendu par « faible poids à la naissance », « faible poids de naissance », « faible poids » ou « (bébés de) petit poids », *high birth weight* par « excès de poids à la naissance » et *unhealty weight* par « poids déficient à la naissance » ou « insuffisance pondérale à la naissance ». Il n'y a pas de différence d'appréhension de la notion (du concept) d'un corpus à l'autre. Les seules différences que l'on observe dans les traductions sont soit d'ordre morpho-syntaxique soit d'ordre discursif dans le cas de la réduction du terme par effacement de l'expansion. Ainsi, à l'exception des variantes discursives, l'ensemble des traductions extraites pour les termes *birth weight* et *low birth weight* (« poids de naissance » et « poids à la naissance » pour le premier, « faible poids à la naissance », « faible poids de naissance » et « insuffisance pondérale à la naissance » pour le second) peuvent être retenues et intégrées dans la terminologie.

	CESART (17)	**Hansard (5)**
low birth weight	*faible poids à la naissance* *faible poids de naissance* *faible poids* *petit poids*	*insuffisance pondérale à la naissance* *insuffisance pondérale* *petit poids*
high birth weigth	*excès de poids à la naissance*	
unhealthy birth weight	*poids déficient à la naissance* *insuffisance pondérale à la naissance*	
unsatisfactory birth weight		(*n'ont*) *pas un poids suffisant à la naissance*

Tableau 8 - Traductions des synonymes birth weight *et* low birth weight

6.3. birth control

Les traductions de *birth control* extraites des corpus touchent soit aux aspects techniques/méthodologiques de la contraception, (« méthodes contraceptives », etc.) soit aux aspects socio-politiques de la natalité (« contrôle des naissances », etc.). Le Tableau 9 présente le détail des traductions extraites pour ce terme selon qu'il apparaisse comme une unité sous-phrastique à part entière ou englobé dans un syntagme nominal plus large (par exemple, *birth control pills* ou *forms of birth control*). La dernière ligne du tableau fait la synthèse des traductions issues de chacun de ces contextes.

	CESART (33)	Hansard (38)
birth control	*méthode(s) contraceptive(s) moyen(s) de contraception contraception*	*contrôle des naissances planification des naissances régulation des naissances méthode(s) de contraception moyen(s) de contraception contraception contraceptifs méthode d'orthogénie*
birth control pills	*pillule anticonceptionnelle contraceptif oral pillule contraceptive*	*pillule anticonceptionnelle*
birth control method(s) / method(s) of birth control	*méthode(s) contraceptives moyen(s) de contraception mode de contraception*	*méthode(s) de contraception méthode(s) de contrôle des naissances contrôler les naissances*
means of birth control		*moyen de limitation des naissances*
measures of birth control		*mesures de contrôle des naissances*
form(s) of birth control	*méthodes de contraception*	*moyen de contrôle des naissances moyen de régulation des naissances*
types of birth control	*modes de contraception*	
birth control device(s)	*moyens de contraception méthodes de contraception*	*moyens anticonceptionnels méthode de contraception*
birth control products		*contraceptifs*
birth control	*méthode(s) contraceptive(s) moyen(s) de contraception contraception anticonceptionnelle contraceptive*	*contrôle des naissances contrôler les naissances planification des naissances régulation des naissances limitation des naissances méthode(s) de contraception moyen(s) de contraception méthode d'orthogénie contraception contraceptifs anticonceptionnelle*

Tableau 9 - Traductions du synonyme birth control

A priori, le corpus Hansard permet d'extraire un plus grand nombre de traductions pour *birth control* (11 traductions, contre 5 pour CESART). Cependant, dans la terminologie MeSH, *birth control* est un synonyme de <*contraception*>, terme dénotant les aspects techniques/méthodologiques du contrôle des naissances. En effet, dans le MeSH, les aspects socio-politiques de la natalité sont dénotés par les termes <*services de planification de la famille*> et <*politique contrôle naissances*>. Près de la moitié des traductions extraites du Hansard correspondent en fait à l'un ou l'autre de ces termes, et non à <*contraception*>. Finalement, « méthode d'orthogénie » est la seule traduction extraite du Hansard et non présente dans CESART que nous pourrons retenir comme synonyme de <*contraception*>.

L'exemple de *birth control* est à rapprocher du cas de *drug cost* exposé plus haut dans la mesure où il met en évidence une différence dans l'appréhension des concepts dans les deux corpus. Comme pour *drug cost*, la vision du Hansard est plus générale, ce qui a pour conséquence positive une plus grande diversité lexicale avec l'emploi de «méthode d'orthogénie » (cf. H4f ci-dessous). Cependant, dans ce cas, l'approche globale de la contraception résulte également en un flou conceptuel avec l'absence de distinction entre les trois concepts *<contraception>*, *<services de planification de la famille>* et *<politique contrôle naissances>*. Ainsi, les emplois de *birth control* dans les phrases H5e et H6e renvoient clairement au même concept : c'est la contraception en tant que technique ou méthode qui est abordée dans les deux cas. Pourtant, ce sont des termes évoquant l'application d'une *<politique [de] contrôle [des] naissances>* qui sont employés dans H5f alors que c'est le concept de *<contraception>* qui est traduit en H6f. Cet usage peut s'expliquer par le fait que la contraception en tant que technique ne survient dans le domaine juridique qu'à travers les lois statuant sur le contrôle des naissances, le fonctionnement des services de planification de la famille, voire de l'avortement. Au contraire, dans le domaine médical, c'est l'aspect technique de la contraception qui prime, et cette notion reste bien différenciée des dispositions légales régentant les différentes pratiques de la contraception.

H4e I do not condone abortion as birth control.

H4f Je n'approuve pas l'avortement en tant que méthode d'orthogénie.

H5e Of the 10 (…) couples in that class (…) six were there because of failed birth control.

H5f Sur les dix couples (…) qui participaient à ce cours, six le suivaient parce que leur méthode de planification des naissances avait échoué.

H6e There must be awareness, among our young people (…), of birth control.

H6f Il faut sensibiliser les jeunes (…) aux méthodes de contraception.

7. Termes spécifiques à l'un ou l'autre corpus

7.1. drug legislation, surgical equipment Les tableaux 10 et 11 présentent les traductions extraites pour deux synonymes présents dans l'un des corpus seulement : il s'agit respectivement de *drug legislation* (présent uniquement dans le Hansard) et de *surgical equipment* (présent dans CESART, mais absent du Hansard).

	CESART (0)	**Hansard (13)**
drug legislation		*loi sur les médicaments mesure législative (portant) sur les médicaments législation sur les médicaments loi sur les drogues loi sur les brevets*

Tableau 10 - Traductions du synonyme drug legislation

Parmi les cinq traductions de *drug legislation* extraites du Hansard, les trois premières peuvent être validées et ajoutées à la terminologie comme synonymes du terme *<législation produit chimique ou pharmaceutique>*. Ces traductions peuvent également permettre d'inférer d'autres variantes possibles du terme, comme « législation portant sur les médicaments ». La quatrième traduction proposée renvoie à un autre sens de *drug*, celui de «drogue » et peut également être ajoutée à la terminologie. Cependant, la dernière traduction « loi sur les brevets » s'éloigne du

concept de <*législation produit chimique ou pharmaceutique*> et ne peut être validée. Cette traduction résulte d'une formulation condensée/réduite renvoyant en fait aux « brevets sur les médicaments » utilisée le contexte spécifique d'énonciation représenté par les phrases H7e/f.

H7e In the case of the patent drug legislation, witnesses were allowed to (…) speak for 40 minutes (…).
H7f Dans le cas de la loi sur les brevets, les témoins avaient été autorisés à parler pendant 40 minutes (…).

La traduction littérale du synonyme *surgical equipment* extraite du Hansard, à savoir « équipement chirurgical » correspond déjà au terme MeSH français, et le valider en tant que synonyme conduirait à créer une redondance. En revanche, les autres traductions peuvent être validées.

	CESART (6)	**Hansard (0)**
Surgical equipment	*matériel chirurgical équipement chirurgical appareil(s) chirurgical(aux)*	

Tableau 11 - Traductions du synonyme surgical equipment

8. Discussion

8.1. Terminologie bilingue et variation

Dans cette étude, nous faisons dès le départ l'hypothèse que la traduction d'un terme est un terme. De fait, nous nous situons dans un contexte de projection d'une terminologie existante en corpus : les unités lexicales que nous traitons ont un statut terminologique établi qui constitue par là-même une donnée a priori. Nous ne cherchons donc pas à trancher sur le statut de terme des traductions extraites, contrairement à d'autres travaux tels que ceux de A. Nazarenko (2002) ou encore M.-C. L'Homme (2004). En effet, ces derniers se situent dans un contexte de construction de terminologies à partir de corpus qui suppose que les unités lexicales traitées n'ont pas de statut terminologique préétabli. Dans le cadre de l'élaboration d'une version du MeSH en chinois mandarin, Lu *et al*. (2005) utilisent une méthode en deux étapes, similaire à celle que nous avons adoptée : la traduction automatique des termes, puis la validation des traductions par des terminologues spécialistes du domaine médical. Dans ces travaux, l'accent est mis sur l'extraction de variantes des termes appartenant au registre du vocabulaire « patient ».

Par ailleurs, pour ce qui est de la variation terminologique en corpus spécialisé, il existe de nombreux travaux portant sur cet aspect. Cependant, la variation terminologique est généralement abordée d'un point de vue monolingue (Daille *et al*., 1996 ; Hamon *et al*. 1998 ; Jacquemin, 1999), les études soulevant cette question d'un point de vue bilingue restent encore marginales (Carl *et al*., 2004). Ainsi, Carl *et al*. (2004) proposent un système de détection semi-automatique de termes et leurs variantes dans des textes parallèles anglais/français. Pour ce faire, ils utilisent un dictionnaire bilingue construit à la main contenant des termes de base et leur traduction ainsi que des patrons de variation terminologique. Les auteurs évaluent le système en fonction des différents patrons de variation pris en compte. Bien que deux corpus aient été utilisés pour

l'évaluation, il s'agit de textes d'un même domaine, militaire, et d'un même type, manuel de formation. En effet, l'objectif premier de cette étude est de mesurer les performances du système en fonction des patrons appliqués. On ne peut donc que regretter l'absence d'une évaluation sur des corpus variés qui aurait permis d'analyser l'évolution des performances en fonction du type de corpus et de statuer sur un lien éventuel entre type de corpus et variation terminologique.

Notre contribution consiste à étudier la variation terminologique dans le domaine médical offerte par des corpus issus de domaines différents. A travers cette étude contrastée, nous cherchons à évaluer l'apport de ressources hors du cœur de la spécialité pour l'enrichissement de terminologies spécialisées. Ici, la spécialité étudiée étant la santé, et les ressources « autres » relevant du domaine général/législatif (Hansard).

8.2. Les traductions extraites pour un même terme sont elles complémentaires ?

La projection de la terminologie MeSH dans les deux corpus d'étude révèle une double complémentarité. Tout d'abord, les termes dont les corpus font usage sont fonction de leur degré de spécialisation dans le domaine médical. L'utilisation d'un corpus non spécialisé comme le Hansard pour l'enrichissement d'une terminologie médicale peut sembler à première vue critiquable. Elle nous a cependant dès le départ paru pertinente en regard de la diversité des catégories des termes représentées dans le MeSH, pertinence confirmée par le fait que certains termes n'apparaissent que dans l'un ou l'autre des deux corpus et qu'il existe par ailleurs un noyau de termes commun aux deux. Ce noyau commun est à l'origine d'une complémentarité sur le plan des traductions trouvées pour les termes anglais dont le résultat est un enrichissement de la terminologie plus important que si un seul corpus avait été utilisé. C'est le cas par exemple du terme *drug cost* pour lequel seul le Hansard propose la traduction « frais pharmaceutiques » en plus des traductions communes aux deux corpus. C'est également le cas du terme *low birth weight* pour lequel chaque corpus propose une/des traduction(s) distincte(s) : « faible poids de naissance », « faible poids à la naissance » pour CESART et « insuffisance pondérale à la naissance » pour le Hansard.

Cette complémentarité tient également au fait que le corpus spécialisé CESART présente une plus grande régularité terminologique que le corpus non spécialisé qu'est le Hansard. Dans celui-ci, le nombre moyen de traductions extraites par terme est plus élevé (cf. Tableau 6) et le corpus offre une plus grande diversité lexicale. Bien qu'une telle supposition demande encore à être validée, cette observation suggère néanmoins qu'il existerait un lien entre niveau de spécialisation d'un corpus et variation terminologique : la variation terminologique serait inversement proportionnelle au degré de spécialisation d'un corpus. Autrement dit, la variation terminologique serait moins fréquente dans un corpus spécialisé que dans un corpus non spécialisé, comme le remarquent notamment Chodkiewicz *et al*. (2002) : « There is [...] a link between formal regularity and subject domain ».

8.3. Les usages renvoient-ils à différents concepts dénotés par les mêmes termes ?

L'étude d'exemples précis présentée ci-dessus met en évidence des différences dans l'appréhension des concepts dans les deux corpus, en particulier pour les termes *drug cost* et *birth control*. Dans le Hansard, ces deux notions sont abordées d'un point de vue plus large et surtout, dans des contextes variés incluant les aspects sociaux, économiques et politiques, qui dans nos exemples priment sur l'aspect strictement clinique. En revanche, dans CESART, le point de vue clinique prime dans la majorité des contextes.

Du point de vue de la terminologie MeSH, on peut donc dire que les usages des termes étudiés peuvent renvoyer à différents concepts dans les deux corpus – par exemple, *birth control* est employé dans le Hansard avec, tour à tour, le sens de <*contrôle des naissances*> ou <*planification des naissances*>, alors que dans CESART, il est toujours employé avec le sens de <*contrôle des naissances*>. Ces différences semblent s'expliquer par la différence de coloration des contextes dans lesquels ils apparaissent, le prisme du Hansard semblant a priori plus large que celui de CESART. Cependant, on peut également s'interroger sur l'influence éventuelle du degré de spécialisation des émetteurs (auteurs des textes étudiés). En effet, bien que la cible des deux corpus soit le grand public, ce qui se manifeste par une sous-représentation des termes techniques très spécialisés, on observe dans CESART une plus grande régularité dans le choix des formulations employées, et de ce fait, moins d'écarts avec les concepts désignés par ces formulations dans le MeSH.

8.4. L'ensemble des usages qui se dégagent des différents contextes doivent-ils être pris en compte dans la terminologie ?

Bien qu'il y ait parfois changement de point de vue, différentes variantes peuvent être intégrées dans la terminologie MeSH. Ainsi, « frais pharmaceutiques » qui est une désignation plus globale par rapport à « coût/prix des médicaments » peut être validé comme synonyme de <*coût médicaments*>. Il n'en va pourtant pas toujours de même, même si la traduction en question est valide. Ainsi, les traductions du terme *birth control* qui renvoient, dans le Hansard, à des aspects socio-politiques de la natalité comme <*contrôle des naissances*> ou <*planification des naissances*> ne peuvent être retenues dans la mesure où c'est l'aspect technique qui est véhiculé par *birth control* dans le MeSH. L'usage du terme tel qu'il est fait dans chaque corpus renvoie à un aspect différent du concept qui, dans le cas du Hansard, n'est pas en adéquation avec la structure conceptuelle de la terminologie. La validité des traductions identifiées dans les corpus est une chose, leur intégration dans la ressource en est une autre : elle suppose une connaissance approfondie de cette dernière et rend par conséquent l'intervention d'un terminologue ou spécialiste du domaine indispensable.

Le travail présenté ici constitue une étude de cas sur la traduction de termes du domaine médical dans deux corpus. Afin de confirmer les observations sur la projection de terminologies spécialisées renvoyée par différents corpus telle que nous l'avons effectuée, il conviendra de multiplier les expériences afin d'élargir le champ d'étude à d'autres termes du domaine médical puis à d'autres domaines de spécialité.

9. Conclusion

Cette étude des termes du domaine médical dans un corpus spécialisé et dans un corpus non spécialisé indique qu'il n'y a pas de séparation nette entre langue générale et langue de spécialité, quels que soient les contextes d'emploi des termes. Nous avons observé un *continuum* entre ces différentes langues. La prise en compte du spectre complet d'un domaine de spécialité - ici, le domaine médical – à différents niveaux du spectre permet de compléter une terminologie du domaine. Dans notre cas, les contextes cliniques d'emploi des termes (issus du corpus spécialisé) aussi bien que les contextes juridiques ou économiques et sociaux (issus du corpus non spécialisé) ont été une source d'apports à la terminologie. La phase de validation par un expert du domaine est néanmoins essentielle pour assurer la cohésion des apports. Ainsi, la variété des corpus contribue à l'enrichissement terminologique mais requiert une grande vigilance afin de

bien identifier les dérives de sens susceptibles d'apparaître à mesure que la coloration du corpus s'éloigne du cœur de la spécialité étudiée.

Remerciements

Ce travail a été réalisé dans le cadre du projet VUMeF, qui bénéficie d'un financement du Réseau National Technologies pour la Santé (RNTS). Nous remercions Philippe Langlais du laboratoire de Recherche appliquée en linguistique informatique de l'Université de Montréal (http://rali.iro.umontreal.ca) de nous avoir fourni le corpus Hansard aligné au niveau des phrases. Nous tenons également à remercier Benoît Thirion, conservateur de la bibliothèque médicale du CHU de Rouen, pour son expertise lors de la validation des traductions.

Bibliographie

CARL M., RASCU E., HALLER J., LANGLAIS Ph., 2004, « Abducing term variant translations in aligned texts », dans *Terminology* 10(1), pp. 103-133.

CHODKIEWICZ C., BOURIGAULT D., HUMBLEY J., 2002, « Making a workable glossary out of a specialised corpus. Term extraction and expert knowledge », dans B. Altenberg and S. Granger (éds.), *Lexis in Contrast. Coprus-based approaches*, John Benjamins, Amsterdam/Philadelphia, pp. 249-267

DAILLE B., 1994, *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*, Thèse de doctorat en Informatique fondamentale, Université Paris 7.

DAILLE B., HABERT B., JACQUEMIN Ch., ROYAUTÉ J., 1996, « Empirical observation of terms variations and principles for their description », dans *Terminology* 3(2), pp. 197-257.

DARMONI SJ., JAROUSSE É., ZWEIGENBAUM P., LE BEUX P., NAMER F., BAUD R., JOUBERT M., VALLÉE H., COTE RA., BUEMI A., BOURIGAULT D., RECOURCÉ G., JENNEAUS., RODRIGUES JM., 2003, « VUMeF: Extending the French Involvement in the UMLS Metathesaurus », dans *Actes de AMIA Symp. 2003*, pp. 824.

GAUSSIER É., 2001, « General considerations on bilingual terminology extraction », dans D. Bourigault, Ch. Jacquemin, M.-C. L'Homme (éds.), *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam/Philadelphia, pp. 167-183.

HABERT, B., NAZARENKO, A., SALEM, A., 1997, *Les linguistiques de corpus*, Paris, Armand Colin/Masson.

HAMON T., NAZARENKO A., GROS C., 1998, « A step towards the detection of semantic variants of terms in technical documents », dans *Actes ACL Conference*, pp. 498-504.

HAMON T., 2000, *Variation sémantique en corpus spécialisé : Acquisition de relations de synonymie à partir de ressources lexicales*, thèse de doctorat en informatique, Université Paris-Nord.

JACQUEMIN Ch., 1999, « Syntagmatic and Paradigmatic representations of term variation », dans *Actes ACL Conference*, pp. 341-348.

L'HOMME M.-C., 2004, « Sélection des termes dans un corpus d'informatique : comparaison des critères lexico-sémantiques », dans *Actes du congrès Euralex*, pp. 583-593.

LU W. H., LIN S. J., CHAN Y. C. and CHEN K. H., 2005, « Semi-automatic construction of the

Chinese-English MeSH using web-based term translation method », dans *Actes AMIA Symp.*, pp. 475-9.

NAZARENKO A., 2002, *Acquisition de connaissances à partir de corpus : élaborer des ressources pour l'extraction d'information*, Rencontres Sémantique et Corpus de l'ERSS.

NÉVÉOL A., OZDOWSKA S., 2005, « Extraction bilingue de termes médicaux dans un corpus parallèle », dans *Actes des 5ème journées Extraction et Gestion des Connaissances* pp. 655-666. Toulouse : Cépadués.

OZDOWSKA S., 2004, « Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés », dans *Actes de RECITAL 2004*, pp. 125-134.

OZDOWSKA S., NEVEOL A., THIRION B., 2005, « Traduction compositionnelle automatique de bitermes dans des corpus anglais/français alignés », dans *Actes de TIA 2005*, pp. 83-94.

ZWEIGENBAUM P., BAUD R., BURGUN A., NAMER F., JAROUSSE É., GRABAR N., RUCH P., LE DUFF F., THIRION B., DARMONI SJ., 2003, « UMLF : construction d'un lexique médical francophone unifié », dans *Actes des JFIM*.

FAIRE SE RENCONTRER LES PARALLÈLES : REGARDS CROISES SUR L'ACQUISITION LEXICALE MONOLINGUE ET MULTILINGUE

Pierre Zweigenbaum

Inserm, U729 ; Inalco, CRIM ; AP-HP, STIM

Benoît Habert

CNRS, LIMSI ; Université Paris X-Nanterre

1. Introduction

L'*acquisition (sémantique) lexicale* (semi-)automatique a pour objectif de constituer ou d'accroître des dictionnaires sémantiques. En contexte monolingue, il s'agit de chercher des relations sémantiques, et en particulier, ce sur quoi nous nous centrerons, de partitionner les mots d'un corpus en classes. Chaque classe, dans l'idéal, rassemble des mots de sens proches : synonymes, antonymes, couples hyponyme-hyperonyme, etc. En contexte multilingue, il s'agit essentiellement de repérer des équivalents traductionnels.

L'*alignement* (multilingue) (Véronis, 2000a,b) part de deux textes qui sont en rapport de traduction. Il consiste à établir des correspondances de plus en plus fines : entre les grandes parties du texte (alignement macro-structurel) ; entre phrases (alignement phrastique) ; entre mots (alignement lexical). Il fournit aux traducteurs professionnels mais aussi aux lexicographes des équivalences traductionnelles qui complètent les usuels existants (en particulier dans les domaines spécialisés). On y trouve des correspondances pour les néologismes, mais aussi des variantes admises pour la traduction d'un mot donné ainsi que des indications sur la traduction la plus adéquate pour un terme (celle qui est homologuée par l'usage). Il existe toutefois un risque de calques et plus généralement de « biais de traduction », comme l'anglais « consistent » (= « cohérent ») traduit en français par « consistant ». L'alignement bénéficie de la multiplicité des ressources en rapport de traduction mutuelle (modes d'emploi, textes officiels dans des communautés ayant plusieurs langues officielles comme le Canada ou l'Europe). Le web constitue à l'évidence un réservoir de textes alignables. Un certain nombre de dispositifs expérimentaux visent à découvrir de tels textes (Resnik & Smith, 2003).

Comme il n'existe pas forcément de corpus alignés disponibles ou rassemblables dans un domaine d'application donné, on peut également constituer des *corpus comparables*. Il s'agit d'ensembles de textes dans deux langues qui ne sont pas en rapport de traduction mutuelle (ce qui rend moins probables les calques) mais qui traitent du même domaine, plus ou moins étroit, et qui relèvent, si possible, du même registre ou genre linguistique. Là encore, le web offre des ressources appréciables. Dans le domaine médical, il est ainsi possible de rassembler des documents ressortissant à la même spécialité (par exemple, la médecine coronarienne) et au même registre (cours d'université). Les corpus comparables pallient alors le manque de corpus alignés. Ils permettent également de constituer semi-automatiquement des ressources morphologiques ou lexicales bilingues.

En alignement, les correspondances structurelles accessibles, même sans descendre en deçà de la phrase, fournissent des indices relativement sûrs d'équivalence entre mots des langues concernées, dans un contexte précis. Les mots qui figurent dans les mêmes « bi-fenêtres » ont des chances d'obéir à une certaine proximité sémantique. Cognats (mots identiques ou de graphie proche d'une langue à l'autre, comme « gouvernement » et « government ») et lexiques bilingues peuvent contribuer à l'alignement. Dans le cadre de corpus comparables, de telles bi-fenêtres n'existent pas. Le recours à des lexiques bilingues, de plus ou moins grande couverture et précision, est alors nécessaire pour amorcer la mise en correspondance des autres mots. Dans les deux cas de figure, les rapprochements entre mots s'opèrent sur la base des proximités de contextes d'emploi, dans une optique distributionnelle qui constitue une extension des approches employées pour des corpus monolingues.

Ce parallélisme entre analyse distributionnelle en corpus monolingues et multilingues nous amène à porter un regard croisé sur ces types de travaux. Nous commençons par présenter en section 2 la mise en évidence d'« airs de famille » entre mots en corpus monolingues, puis les adaptations nécessaires pour les corpus comparables (section 3). Les méthodes et enseignements de chacun de ces deux types de travaux devraient pouvoir être réinvestis dans l'autre. C'est ce que nous examinons dans le reste de cet article, d'abord du monolingue vers le multilingue (section 4), puis dans le sens inverse (section 5). Nous concluons sur des considérations d'évaluation et de généralisabilité de ces méthodes (section 6).

2. Contextes et analyse distributionnelle monolingue

L'hypothèse distributionnaliste fondatrice en acquisition sémantique lexicale est que deux mots ont un sens proche s'ils sont employés dans des contextes très voisins. C'est la phrase de Firth (1957), souvent citée : *On reconnaît un mot à ses fréquentations (You shall know a word by the company it keeps*). Harris (1991) en fournit une représentation plus stricte : le fait que deux mots soient opérateurs (gouverneurs) et/ou opérandes (gouvernés) des mêmes ensembles de mots les rapproche[9]. Pour le français, en dehors de domaines de spécialité, c'est l'approche poursuivie par G. Gross dans la définition de *classes d'objets* (Gross, 1994 ; Le Pesant, 1994).

En corpus monolingue, trois grandes étapes (Grefenstette, 1994a) – autorisant chacune de nombreuses variantes – sont mobilisées pour dégager des « airs de famille » entre mots :

1. caractérisation des mots par les contextes dans lesquels ils figurent ;
2. obtention d'un indice synthétique de similarité/distance entre mots ;
3. regroupement des mots selon les distances qui les caractérisent.

[9] Voir (Habert & Zweigenbaum, 2002 : 89-95) pour une présentation détaillée.

La première étape est celle de la représentation des emplois des mots en contexte par des traits jugés pertinents. Le contexte, qui peut être une proposition, une phrase, un paragraphe, un document, est souvent réduit aux mots qui le constituent. Chacun de ces contextes équivaut à une *fenêtre* au sein de laquelle on examine le comportement du mot, en particulier ses rencontres ou *cooccurrences* avec d'autres mots. On se limite souvent aux formes canoniques (*lemmes*) des mots dits « pleins », par opposition aux mots-outils ou mots dits « vides » (déterminants, adverbes, conjonctions). Dans le tableau (a) de la figure 2, se trouvent de telles fenêtres. Ainsi le mot m_1 coocurre avec le mot m_3, ainsi qu'avec le mot m_n dans la fenêtre f_1. Cela permet de nourrir le tableau (b) de la figure 2: il récapitule pour chaque mot (en ligne) les traits (ici, les mots) avec lesquels il cooccurre (en colonnes). Ces simples décomptes de nombres de cooccurrences sont en général remplacés par un indice de force d'association entre le mot et le trait. On peut par exemple pondérer le nombre de cooccurrences d'un mot avec un trait par le nombre de cooccurrences dans lesquelles figure ce mot et par le nombre de mots qui emploient ce trait : en recherche d'information, c'est la famille de pondérations dite TF.IDF – *Term Frequency.Inverse Document Frequency* (Manning & Schütze, 1999).

(a) « Fenêtres » de cooccurrences

f_1		m_3	m_1			m_n	
f_2		m_1	m_4	m_3			
f_3	m_2		m_5		m_n		m_4
f_4			m_2	m_5		m_n	
f_5		m_3			m_2	m_n	
⋮							

(b) Récapitulatif des cooccurrences

	m_1	m_2	m_3	m_4	m_5	⋯	m_n
m_1			2	1			1
m_2			1	1	2		3
m_3	2	1		1			2
m_4	1	1	1		1		1
m_5		2		1			2
⋮							
m_n	1	3	2	1	2		

(c) Ingrédients d'un indice de similarité

		m_1	
		traits présents	traits absents
m_2	traits présents	m_3 m_4 m_n	m_5
	traits absents	...	

(d) Indice de Jaccard

$$\frac{|\text{traits}(m_1) \cap \text{traits}(m_2)|}{|\text{traits}(m_1) \cup \text{traits}(m_2)|} = \frac{3}{4} = 0,75$$

FIG. 1 : Proximités entre mots – corpus monolingues

La deuxième étape consiste à résumer les rapprochements/éloignements entre les mots deux à deux. Elle se base en particulier sur le nombre de traits partagés par les deux mots[10] et sur le nombre de traits propres à chacun d'eux (tableau (c) de la figure 2 pour les mots m_1 et m_2), ainsi que sur la fréquence globale de chaque trait et sur le nombre de traits utilisés par chaque mot. L'indice synthétique de Jaccard – en (d) dans la figure 2 – opère un tel résumé. On constate que cet indice de similarité varie entre 1 quand tous les traits sont partagés par les deux mots et 0 quand les deux mots ne présentent aucune intersection (pour m_1 et m_2, il vaut 0,75 : les 2 mots

[10] Comme indiqué brs de la description de la première étape, chaque trait est généralement représenté par une mesure de sa force d'association avec le mot considéré qui va au-delà d'un simple nombre de cooccurrences.

partagent les trois quarts des traits qu'ils emploient). De multiples indices sont d'ailleurs disponibles (Losee, 1998 : 43-62). La distance entre les mots est l'inverse de la similarité : un mot est d'autant plus proche d'un autre que la similarité entre eux est grande.

La troisième étape consiste à regrouper les mots en sous-ensembles en fonction des distances découlant de l'étape précédente. Le regroupement peut reposer sur la classification hiérarchique ascendante (Lebart *et al.*, 1997 : 155-176) : on commence par regrouper les mots ou les groupes de mots entre lesquels la distance est la plus faible, puis on agrège un mot ou un groupe de mots un peu plus éloigné (la manière de calculer la distance entre un groupe déjà constitué et les mots encore « libres » ou d'autres groupes est un paramètre de la classification) et l'on continue jusqu'à obtenir un arbre de mots ou *dendrogramme*. Pour obtenir un ensemble de «classes » du niveau de finesse souhaité, cet arbre peut être ensuite élagué à un niveau donné (on garde les nœuds de profondeur k) ou en conservant des nœuds de diverses profondeurs selon un critère de qualité des nœuds en question (Jardino, 2004 ; Rossignol, 2005). On peut également obtenir directement un regroupement en ensembles disjoints : ce sont les techniques d'agrégation autour de centres mobiles ou «nuées dynamiques » (*k-means*) (Lebart *et al.*, 1997 : 148-154). Dans les deux cas, le nombre de classes que l'on retient est un paramètre important. Au delà de quelques dizaines voire d'une dizaine de classes, les résultats sont peu compréhensibles et les raisons des distinctions difficiles à comprendre et à expliciter. C'est en outre par commodité et avec optimisme qu'on dénomme *classes* les regroupements résultants. S'ils comprennent effectivement des mots en relation de synonymie, d'hyperonymie/hyponymie, d'antonymie, ils incluent également des relations plus complexes (méronymie ou relation de partie à tout) et des rapprochements plus douteux. Il reste donc toujours à les trier (éliminer les intrus au sein d'un regroupement et enlever les groupes « poubelles ») et ensuite à les interpréter, c'est-à-dire minimalement à leur attribuer une étiquette qui « résume » leur contenu.

3. Analyse distributionnelle sur corpus comparables

3.1. Adaptation au cadre bilingue

Dès lors qu'on ne dispose pas de suffisamment de corpus parallèles, ou que l'on cherche à éviter les biais de traduction, les corpus comparables constituent un réservoir de matériau linguistique qui peut aider à construire ou étendre des ressources lexicales bilingues. Dans ces corpus, la correspondance structurelle systématique qui existait dans les corpus parallèles au niveau des documents, des phrases (dans une large mesure) et des mots (avec de nécessaires ajustements) disparaît. Il faut donc s'appuyer sur d'autres indices de correspondance. Rappelons que l'objectif est de détecter des couples de mots (mot_s, mot_c) des corpus source C_s et cible C_c en relation de traduction[11] : des mots possédant donc un sens proche, mais appartenant à deux langues différentes, représentées par les deux corpus «comparables ». Une méthode amorcée par Rapp (1995)[12] consiste à exploiter l'hypothèse distributionnaliste citée plus haut (« deux mots ont un sens proche s'ils sont employés dans des contextes très voisins »), en l'étendant au cas

[11] Nous employons les termes « source » et « cible » par commodité, sans préjuger de l'ordre dans lequel on utilise les deux ensembles de textes. Par ailleurs, les corpus comparables contiennent quelquefois aussi des textes en rapport de traduction. Nous ne préjugeons alors pas non plus de l'ordre dans lequel ils ont été écrits : source traduite en cible, cible traduite en source, source et cible traductions d'une troisième langue, source et cible corédigées, etc.

[12] Fung & McKeown (1997) font remonter l'idée à Dagan *et al.* (1991), qui ont proposé d'employer des connaissances sur les cooccurrences dans une langue cible pour désambiguïser un mot polysémique d'une langue source.

bilingue. On cherche alors à identifier des couples de mots tels que la distribution du mot source dans le corpus source et la distribution du mot cible dans la langue cible soient similaires. Dans la mesure où la distribution est une caractérisation indirecte du sens, on aura ainsi des mots de sens proches.

Cette distribution sera calculée sur chacun des deux corpus de la façon vue plus haut sur un corpus monolingue : typiquement donc, en repérant dans une fenêtre mobile les traits avec lesquels un mot cooccurre (ses contextes) et en collectant leur décompte dans des vecteurs de contextes (les rangées du tableau b de la figure 2). Les mêmes paramètres sont à fixer : taille de la fenêtre[13], mesure de la force d'association entre mot et trait, etc. Un vecteur représente la distribution d'un mot dans un corpus ; les dimensions du vecteur (les colonnes du tableau) sont la liste des traits pour ce corpus (généralement, la liste des « mots pleins » de ce corpus).

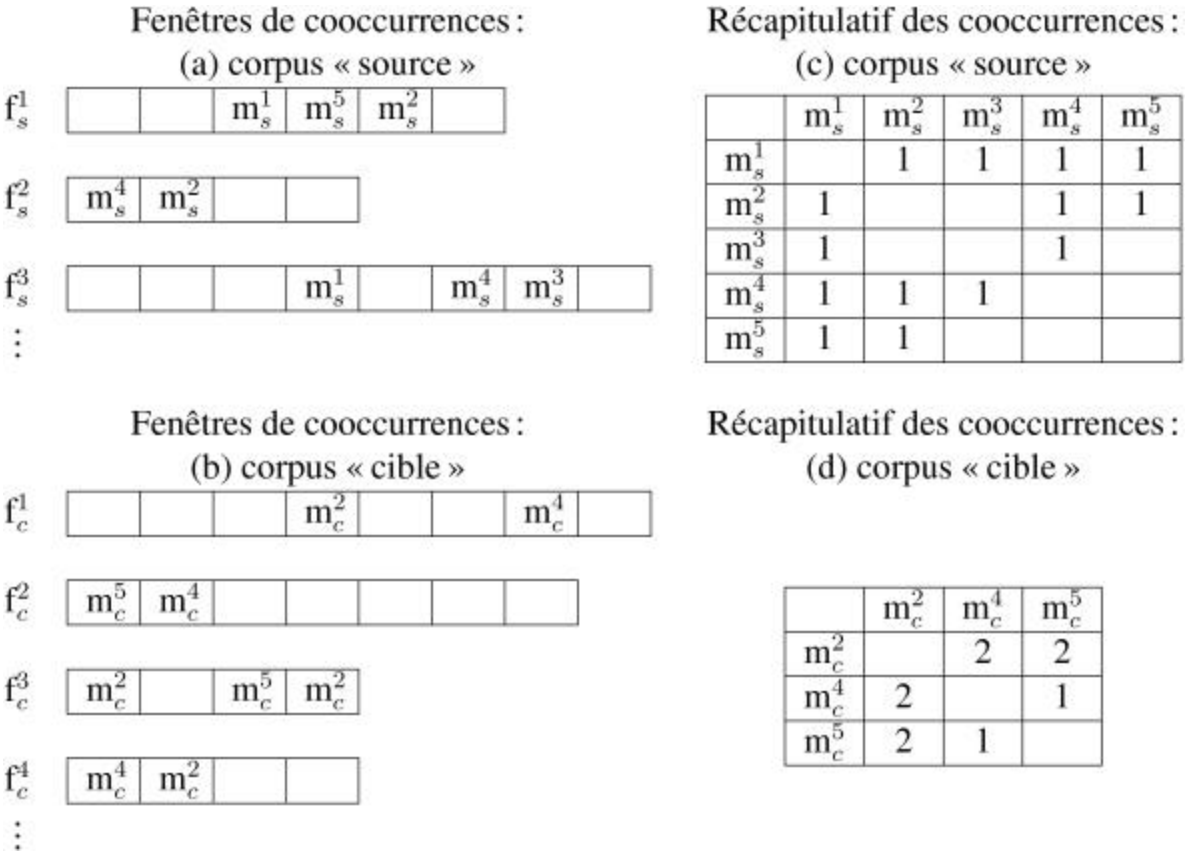

FIG. 2 : Proximités entre mots – corpus comparables (1/2)

Les corpus comparables, comme le manifestent les tableaux (a) et (b) de la figure 2, ne disposent pas du couplage entre fenêtres de même grain qui caractérise les corpus alignés. Ils supposent alors un *lexique pivot* bilingue (tableau e, figure 3) dans le sens langue source-langue cible (ou e^{-1} dans le sens langue cible-langue source) qui permet pour une fenêtre dans la

[13] Une fenêtre de la taille d'une phrase ou moins permet d'approximer des relations syntaxiques entre mots (dépendance entre gouverneur et gouverné). C'est le cas de Sadat *et al.* (2003) et Chiao & Zweigenbaum (2002), qui prennent une fenêtre de deux à trois mots. Si la fenêtre est plus grande que la phrase, on fait appel à un autre type de relation entre mots : une relation d'association thématique (comme les *isotopies*, récurrences sémantiques de Rastier (1987)). C'est ce que font Fung & McKeown (1997), en prenant une fenêtre de la taille d'un paragraphe, pour la recherche de traductions de termes polylexicaux. Déjean *et al.* (2002) utilisent une fenêtre de plusieurs phrases.

première langue d'en produire une « traduction » mot à mot dans l'autre langue. Ces fenêtres traduites permettent de remplir les traits pour l'autre langue et ramènent au cas de figure précédent (figure 2). Le tableau (c'), «traduction » du tableau (c), est comparable avec le tableau (d) et inversement le tableau (d') est comparable avec le tableau (c). On note que ce lexique bilingue ne couvre pas, en général, tous les mots de la première langue : dans le tableau (e), le mot m_s^5 ne comporte pas de traduction[14]. Inversement, un trait dans une langue peut avoir plusieurs équivalents dans l'autre : c'est le cas du mot m_s^3 dans le tableau (e) et de m_c^5 dans le tableau (e^{-1}).

(e) Lexique pivot : source → cible

m_s^1	m_s^2	m_s^3	m_s^4	m_s^5	m_s^n
m_c^2	m_c^4	m_c^2 m_c^5	m_c^5		m_c^1

(c') Traduction du tableau (c)

	m_c^2	m_c^4	m_c^5
m_c^2		1	2
m_c^4	1		1
m_c^5	2	1	

(e^{-1}) Lexique pivot : cible → source

m_c^1	m_c^2	m_c^4	m_c^5
m_s^n	m_s^3	m_s^2	m_s^4 m_s^3

(d') Traduction du tableau (d)

	m_s^2	m_s^3	m_s^4
m_s^2		3	1
m_s^3	3		2
m_s^4	1	2	

FIG. 3 : Proximités entre mots – corpus comparables (2/2)

Les vecteurs de contextes résultants ont donc une dimension qui correspond à l'intersection du lexique d'une part (de son côté source) avec les traits du corpus source, et d'autre part (de son côté cible) avec les traits du corpus cible. En d'autres termes, les mots qui pourront servir de traits communs sont les couples de mots du lexique dont le membre source est présent dans le corpus source et dont le membre cible est présent dans le corpus cible. Les vecteurs de contextes source et cible sont projetés sur ce référentiel commun.

Une fois projetés, les vecteurs de contextes source et cible deviennent comparables. Il est alors possible de déterminer, pour un vecteur de contextes représentant un mot source, quels vecteurs de contextes cible sont les plus similaires (Fung & McKeown, 1997 ; Rapp, 1999 ; Déjean *et al.*, 2002 ; Chiao & Zweigenbaum, 2002). On a ainsi étendu l'analyse distributionnelle aux deux corpus, en la synchronisant à travers le lexique pivot.

[14] On part ici du principe que ce lexique est partiel. En effet, s'il était complet, on disposerait d'une correspondance pour chaque mot du corpus, et le problème initial (trouver de telles correspondances) ne se poserait plus vraiment. En pratique, la recherche d'autres traductions pourrait probablement tout de même s'avérer utile, par exemple pour étendre un lexique existant ou adapter un lexique général à un type de corpus spécifique : textes spécialisés, autre niveau de langue, particularismes régionaux, etc.

3.2. Exemple : structure de deux corpus médicaux comparables

Comme dans tout travail sur corpus, la constitution même du corpus conditionne la réussite du travail entrepris. Comment faire en sorte que deux corpus soient les plus comparables possibles, c'est-à-dire rapprochent des emplois des deux langues les plus similaires ? Nous examinons cette question à travers l'exemple du corpus de travail français-anglais de Chiao & Zweigenbaum (2002), que nous appellerons [C4] (Corpus comparable CISMeF - CliniWeb). Le corpus [C4] concerne le domaine médical. Pour obtenir un corpus comparable, il fallait s'assurer que le domaine couvert par chacune des parties était semblable. Pour cela, Chiao & Zweigenbaum se sont appuyés sur l'existence de catalogues de sites web médicaux : pour le français, CISMeF (Catalogue et index des sites médicaux francophones) (Darmoni *et al.*, 2000)[15] et pour l'anglais, CliniWeb (Hersh *et al.*, 1999) (site fermé depuis). CISMeF recense les sites web médicaux francophones de qualité, et les indexe par des mots clés pris dans une terminologie basée sur le thésaurus hiérarchique MeSH[16] (Medical Subject Headings). On notera, malgré la proximité de domaine, certifiée par une indexation commune, le fort décalage de taille entre les deux parties de ce corpus (de 1 jusqu'à 15 selon la version) qui nuit probablement à la comparaison. Notons que des différences importantes entre corpus source et cible sont aussi observées dans d'autres travaux : Sadat *et al.* (2003) mettent en correspondance 13,5 Mmots japonais avec 1,5 Mmots anglais.

L'un des intérêts de l'indexation contrôlée, réalisée par des documentalistes médicaux, est qu'elle permet de sélectionner un ensemble de sites consacrés à un sous-domaine déterminé. De façon similaire, CliniWeb recensait des sites médicaux anglophones, en les indexant par des mots clés du thésaurus MeSH. Dans la première version du corpus [C4], Chiao & Zweigenbaum ont choisi de travailler sur des pages web parlant de *signes et symptômes* (catégorie MeSH *C23* : corpus [C4-23]). Ils ont pour cela extrait de ces catalogues les adresses (*URL*) des pages catégorisées par cette catégorie ou l'un de ses descendants[17]. Ils ont ensuite téléchargé ces pages web, et les ont converties de HTML ou PDF en texte brut. Cela a donné un corpus composé d'une partie obtenue à travers CISMeF et parlant de signes et symptômes en français (10 539 fichiers, 16,7 Mmots), et d'une partie obtenue à travers CliniWeb et parlant de signes et symptômes en anglais (2 036 fichiers, 1,1 Mmots).

Dans une seconde version du corpus, [C4-tout], ont été téléchargées l'ensemble des pages pointées par CISMeF et l'ensemble des pages pointées par CliniWeb (avec le même ajustement sur les pages immédiatement inférieures que pour [C4-C23]). Au total, [C4-tout] contient 32 951 fichiers et 54,5 Mmots en français et 11 755 fichiers et 7,6 Mmots en anglais. Il a ainsi été possible de travailler aussi bien à un niveau sous-domanial ([C4-C23], signes et symptômes) que domanial ([C4-tout], santé).

Dans l'idéal, les différentes dimensions qui caractérisent un corpus (voir par exemple Habert *et al.* (2001)) devraient être maîtrisées de la même façon dans les deux parties du corpus comparable. Nous développons ces questions dans la section suivante.

[15] http://www.chu-rouen.fr/cismef/
[16] http://www.nlm.nih.gov/mesh/meshhome.html
[17] Pour diverses raisons expliquées dans (Chiao & Zweigenbaum, 2002), les URL de CISMeF pointent quelquefois sur des pages parentes de la page d'intérêt ; ont donc été aussi systématiquement téléchargées les pages immédiatement inférieures à la page pointée, en restant sur le même site.

4. Enseignements de l'analyse de corpus monolingue

L'histoire plus longue de l'acquisition lexicale monolingue a amené à identifier deux types de paramètres : l'espace de travail que l'on se donne et les modes de représentation du contexte.

4.1 Maîtriser l'espace de travail : domaines, genres, spécialisation, etc.

Constituer des corpus pour l'acquisition lexicale, c'est (chercher à) maîtriser plusieurs axes de variation au sein des contextes constitués pour les mots à organiser : la thématique (ou domaine) d'une part, le genre ou registre d'autre part.

4.1.1 Partitionner en domaines

Pour Rastier *et al.* (1994), une partie des phénomènes de polysémie, qui viennent contrecarrer les approches classificatoires de la section 2, sont artefactuels : ils proviennent de la rencontre artificielle, « organisée » entre des mots qui « habitent » des usages autrement sans partage naturel. Le barrage hydraulique comme ouvrage d'art et le barrage policier ou militaire (Véronis, 2004 ; Ferret, 2004) relèvent de thématiques ou domaines largement disjoints (dans un journal comme *Le Monde*, la rubrique économique d'un côté, l'actualité politique nationale ou internationale de l'autre). La conséquence logique de cette analyse est d'opérer en amont une répartition en domaines des documents à utiliser.

Cette répartition peut résulter d'une classification automatique. C'est la démarche de Rossignol (2005), qui prolonge celle de Pichon & Sébillot (1999). Le corpus utilisé, monolingue, rassemble 14 ans de la partie proprement journalistique (par opposition au courrier des lecteurs, par exemple) des archives du mensuel *Le Monde diplomatique* (1985-1998) : 5 704 articles, dont sont retenus les 98 432 paragraphes de plus de 20 mots, pour qu'ils soient « classables » (soit 11 380 197 occurrences). Le système FAESTOS vise à répartir ce corpus en sous-corpus thématiques, chaque thème rassemblant des textes dont les mots relèvent d'un domaine donné. Il s'agit d'un apprentissage non supervisé et non de l'affectation d'une entité textuelle à un thème choisi dans un ensemble prédéfini (apprentissage supervisé). Le paragraphe est alors considéré comme l'*unité textuelle «atomique », à l'échelle de laquelle se développent les phénomènes d'isotopie* (Rossignol, 2005 : 47), c'est-à-dire de partage ou de convergence sémantique entre mots. Il est compris comme un « sac » de mots (au sens de la recherche d'information). Une variante de classification hiérarchique ascendante est mobilisée pour organiser en arbres les noms et adjectifs (en fonction des paragraphes où ils apparaissent) ainsi que les paragraphes (en fonction des mots qu'ils emploient). Un critère de qualité isole dans un deuxième temps les classes de mots qui sont confirmées par les classes de paragraphes. Ce critère de qualité sert ensuite à la réorganisation des arbres de classes de mots : une fusion entre deux classes n'est retenue que si elle fait progresser ce critère. Il permet aussi des réaffectations.

Le résultat est un ensemble de classes disjointes issues de l'arbre de classification mais différentes des simples coupes à un niveau donné de ce dernier. Un paragraphe est affecté à un thème s'il comporte au moins 2 mots-clés de la classe correspondante. Sert de nom à une classe le *sous-ensemble de trois mots tel que l'ensemble des paragraphes contenant au moins l'un de ces mots inclue une partie la plus étendue possible de l'ensemble des paragraphes* [affectés à cette classe]. Par exemple < enseignement/école/université > et < producteur/agriculteur/céréale >. Un paragraphe peut être affecté à plusieurs thèmes.

Est opéré dans une dernière phase le découpage du corpus de départ en sous-corpus thématiques allant de quelques dizaines à quelques centaines de milliers de mots. Ces sous-

corpus thématiques ne constituent pas une partition puisqu'ils ne sont pas disjoints : un paragraphe peut relever de plusieurs thèmes. M. Rossignol constate d'ailleurs *la proportion relativement élevée de paragraphes reconnus comme abordant plusieurs thèmes, qui représentent environ 36% des paragraphes couverts : on compte en moyenne 1,5 thème par paragraphe, le maximum étant atteint par un long paragraphe détecté comme développant huit thèmes distincts (et les abordant en effet, comme un contrôle manuel a pu le confirmer)* (p. 81).

Deux enseignements peuvent être retirés de l'approche de Rossignol (2005). En premier lieu, les textes concrets font se rencontrer les domaines, d'où la nécessité de pouvoir affecter la fenêtre textuelle choisie à plusieurs thèmes. Cette nécessité vaut pour le paragraphe. Elle demeure *a fortiori* pour le document, mais elle s'avère probablement valide, souvent, pour la phrase. En second lieu, l'éclatement en sous-corpus d'un corpus de départ qui multipliait les polysémies artificielles est un cadeau embarrassant. Les parties non disjointes obtenues présentent certes moins de polysémies. Mais elles sont en même temps de tailles inégales et certaines sont petites. Il faut donc mettre en place des stratégies de compensation du faible nombre de contextes d'emploi qu'elles offrent pour les mots qui y figurent.

4.1.2 Tenir compte des registres (genres)

Biber utilise les divisions d'un corpus de référence en registres grossiers pour montrer que la probabilité d'apparition d'une catégorie morpho-syntaxique donnée est fonction du genre (tableau 1). Dans le corpus LOB, équivalent britannique du corpus Brown[18] pour l'anglais, la probabilité pour certains mots d'appartenir à telle ou telle catégorie change significativement selon qu'on a affaire à des textes de fiction ou à des textes qui ne relèvent pas de la fiction. Le constat est particulièrement frappant pour les mots grammaticaux, dont on s'attendrait à ce qu'ils ne soient pas affectés par le changement de genre textuel. Biber indique en outre que les séquences de probabilités de catégories morpho-syntaxiques (bigrammes), tout comme les collocations, varient également avec le domaine.

[18] En 1979, est rendu librement accessible un ensemble d'un million de mots, le corpus Brown (http://helmer.aksis.uib.no/icame/brown/bcm.html). Sa conception repose sur l'hypothèse variationniste suivant laquelle l'usage d'une langue change selon qu'il s'agit de l'écrit ou de l'oral, et pour chacune de ces grandes dimensions, selon les situations de communication, le domaine, etc., ce qui est souvent résumé sous le chapeau flou de *genre*. Ce corpus rassemble donc 500 extraits de 2000 mots chacun, provenant de textes américains publiés en 1961 et relevant de 15 « genres » (reportage, écrits scientifiques et techniques, etc.). Ce corpus est étiqueté : chaque mot est muni d'une étiquette morpho-syntaxique (partie du discours et précisions).

Forme	cat.	fiction %	non fiction %
trust	N	18	85
	V	82	15
rule	N	31	91
	V	69	9
major	titre	69	11
	A	31	85
that	dét.	37	17
	conj.	45	69
	rel.	14	11
before	prép.	30	54
	conj.	48	32
	adv.	22	14

Tableau 1 : Probabilité d'apparition d'une catégorie morpho-syntaxique donnée est fonction du genre (fiction/ non fiction)

D. Biber a plus généralement cherché à décrire de manière empirique les régularités linguistiques présidant à l'organisation des énoncés en « genres » ou « registres » (*registers*), c'est-à-dire en emplois du langage définis situationnellement et fonctionnellement. Une première étape (Biber, 1988) a consisté à faire émerger des constellations de traits linguistiques, appelées *dimensions* par Biber et mobilisées diversement selon les registres. Pour ce faire, Biber a étiqueté de manière semi-automatique (avec vérification manuelle) la présence de 67 traits linguistiques (marqueurs de temps, d'aspect, pronoms et pro-verbes, passifs, modaux...) dans des extraits de 1 000 mots prélevés dans 481 textes d'anglais contemporain écrit et oral. Ces textes provenaient de deux corpus «panachés » – dans la tradition des corpus anglo-saxons dits «représentatifs » –, et organisés en registres : le corpus LOB (décrit plus haut : 1 000 000 mots en 15 genres de prose « informationnelle » *vs*. de fiction) et le corpus London-Lund (485 000 mots d'anglais parlé : conversations, cours, émissions radiophoniques. . .). L'hypothèse sous-jacente est que certains traits ont tendance à apparaître ensemble, de manière fréquente, et que dans le même temps d'autres traits sont évités.

C'est une telle convergence dans l'emploi de certains traits et l'évitement d'autres et son interprétation que Biber appelle une *dimension*. Les outils de la statistique multidimensionnelle permettent en effet de dégager automatiquement ces attirances et ces rejets. L'examen d'un regroupement opéré et le retour aux textes qui l'exemplifient particulièrement permettent dans un deuxième temps d'interpréter la constellation de traits induite. Par exemple, Biber dénomme *orientation narrative* les emplois convergents et fréquents de verbes au passé, de pronoms à la 3ème personne, de verbes dits «publics » (*to complain*), de propositions participes, qu'il oppose à une *orientation non narrative*, caractérisée par la forte présente conjointe de noms, de mots longs (en nombre de caractères), de prépositions, d'adverbes de lieu. Il ne s'agit pas de caractériser en tant que tels des types de textes postulés, mais de partir d'un premier regroupement, lâche, en registres, pour déboucher sur des types de textes. En outre, l'hypothèse n'est pas celle de *marques*, qui seraient associées de manière presque bi-univoque à des types ou à des registres, mais de dimensions (c'est-à-dire de corrélations fonctionnellement cohérentes de sur-emplois et de sous-emplois de traits linguistiques déterminés), qui sont plus ou moins sollicitées par les textes. Les dimensions dégagées par Biber (production impliquée *vs*. informationnelle ;

orientation narrative *vs*. non narrative ; référence explicite *vs*. dépendant de la situation d'énonciation ; visée persuasive explicite ; style abstrait) permettent en effet dans un deuxième temps de regrouper automatiquement les textes qui utilisent ces dimensions de la même manière (en employant préférentiellement un sous-ensemble de dimensions et en évitant un autre sous-ensemble). Biber obtient alors ce qu'il appelle des types de textes : interaction interpersonnelle intime, interaction informationnelle, exposé « scientifique », exposé savant, fiction narrative, récit, reportage situé, argumentation impliquée. Biber caractérise ensuite les genres par les dimensions qu'ils privilégient et par les types de textes qui y dominent.

La deuxième étape du travail (Biber, 1995) a consisté à examiner la généralisation possible de ces dimensions sous-jacentes à d'autres langues que l'anglais, en l'occurrence le coréen, le somali et le nukulaelae tuvulan (langue parlée par 350 personnes sur l'atoll Nukulaelae du groupe Tuvalu dans le Pacifique). La généralisation est à l'évidence contrariée par les écarts de statut entre les langues retenues et les conditions de leur étude (*literacy*, disponibilité de corpus et d'outils de traitement, etc.). Les traits linguistiques associés à une dimension partagée peuvent varier d'une langue à l'autre (ce qui conforte l'hypothèse de dimensions plutôt que de marques). Certaines dimensions sont propres à une ou plusieurs langue(s) étudiée(s) : *honorifics* et *self-humbling* n'existent qu'en coréen. Malgré ces obstacles, Biber estime que les dimensions mises en évidence par la première étape se trouvent confirmées globalement par les recherches de la seconde.

Si l'on réexamine à cette aune le travail de Chiao & Zweigenbaum (2002), on constate qu'ils ont raisonné essentiellement en termes de domaines : ils ont contrôlé le thème des documents, mais pas nécessairement leur genre. Certes, la constitution des deux catalogues exploités, CISMeF et CliniWeb, qui sont consacrés à l'indexation de documents diffusant sur le web des informations et des connaissances en santé, limite de fait les genres présents dans ces corpus. Mais, comme l'indiquent les *types de ressources* que CISMeF assigne à chaque document indexé, la partie française du corpus comprend néanmoins un nombre important de genres textuels différents : guides de bonnes pratiques à l'usage des médecins, sites associatifs à destination des patients, cours et polycopiés à l'intention des étudiants en sont quelques exemples. La partie anglaise du corpus contient elle aussi une variation importante sur ce plan, dans des proportions qui n'ont a priori aucune raison d'être identiques. Il est ainsi probable que des usages différents des mêmes termes aient été mélangés, menant à des vecteurs de contextes moins pertinents : moyennes de contextes reflétant des usages différents, avec des représentations de ces usages variables dans les deux corpus.

4.1.3 Savoir de quoi un corpus est représentatif

Le contrôle des deux dimensions de variation qui viennent d'être détaillées est fondamental pour un usage raisonné des ressources textuelles disponibles pour constituer des corpus comparables, par exemple sur le web. Les conséquences d'une absence de contrôle sont difficiles à prédire. Une différence de genre d'une langue à l'autre, entre les textes dans lesquels un mot est employé, va-t-elle conduire à une mauvaise proposition d'équivalents traductionnels, ou bien va-t-elle être « gommée » et mener tout de même à une proposition correcte ? Pourrait-on trouver des exemples de mots qui risqueraient d'être mal mis en correspondance du fait qu'ils seraient employés majoritairement dans deux genres différents dans les deux corpus ? Des expérimentations ciblées pourraient chercher à mettre en évidence ce type d'effet en construisant des corpus volontairement composés de panachages différents de genres textuels.

Concomitamment, comme l'indiquent Kilgarriff & Grefenstette (2003), le profilage de textes (Illouz *et al.*, 1999 ; Karlgren, 1999) constitue une nécessité pour remplacer la problématique des

années 1970-1990 : *comment créer des corpus représentatifs ?* par la capacité à dire de quels usages spécifiques sont représentatives les données textuelles que l'on peut rassembler pour un besoin déterminé, ici l'acquisition lexicale, monolingue ou multilingue. Il s'agit de coupler une caractérisation fine des documents à tous les niveaux de l'analyse linguistique, à l'aide des instruments dont on dispose désormais (étiqueteurs, parseurs robustes, etc.), et une connaissance détaillée de leur ancrage situationnel et fonctionnel. Est alors crucial le développement de métadonnées détaillées (Habert, 2005 : chapitre VIII), telles que celles postulées par la TEI (*Text Encoding Initiative* – http://www.tei-c.org/) dans son cartouche (*header*) ou par OLAC (*Open Language Archives Community* – http://www.language-archives.org/), s'appuyant sur la proposition du Dublin Core (http://dublincore.org/).

4.2. Ne pas se cantonner aux traits syntaxiques

Le début de la section 2 présentait 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 (Grefenstette, 1994b ; Bourigault, 2002 ; Lin & Pantel, 2002).

Le choix entre les deux pôles est souvent vécu et présenté comme le simple fruit de la nécessité. Si l'on dispose d'un analyseur syntaxique, alors on utilise les dépendances qu'il fournit, sinon on se rabat sur des lemmes étiquetés, voire sur des mots « bruts » ou racinisés[19]. Ce serait une version de l'adage : faire de pauvreté vertu. Toutefois, les deux pôles sont connotés, implicitement, de manière opposée. Les dépendances syntaxiques seraient « justes », elles offriraient une image véridique des contextes des mots. Les lemmes ou les mots bruts en constitueraient une version approchée et, pour tout dire, dégradée.

4.2.1 Pauvretés de la vertu

Cette valorisation a priori est relativisée par la prise en compte du volume global de texte utilisable. Grefenstette (1996) compare précisément les résultats obtenus sur un même corpus avec les deux types de contextes. Dans l'approche syntaxique, les contextes d'un nom sont constitués par les adjectifs, les noms et les verbes avec lesquels il rentre dans une relation de dépendance (en position de gouverneur ou de dépendant). Les relations de dépendance sont fournies par l'analyseur robuste que Grefenstette a développé : Sextant (Grefenstette, 1994b). Dans l'approche « pauvre », les contextes d'un nom sont représentés par tous les noms, tous les adjectifs et tous les verbes dans les dix mots avant ou après, et au sein de la même phrase. La pauvreté est donc relative puisque les mots sont déjà étiquetés et lemmatisés. La mesure de distance est celle de Jaccard (pondérée : le nombre d'occurrences de chaque contexte est pris en compte). Grefenstette utilise comme corpus des phrases de l'encyclopédie *Grolier* contenant un des trente hyponymes du mot institution (comme *establishment*, *charity* ...) dans le dictionnaire sémantique *WordNet*. Le corpus dépasse les 400 000 mots, soit la taille de quatre romans de taille moyenne. Pour pouvoir comparer les deux approches, Grefenstette prend le thésaurus *Roget* comme pierre de touche. Pour un mot donné et une approche donnée, il regarde si son plus proche voisin (le mot avec lequel la proximité est la plus forte selon l'indice de Jaccard) relève de

[19] Nous employons ici *racinisation* comme traduction usuelle mais pas entièrement juste de l'anglais *stemming*, pour désigner la tentative de réduction d'un mot à sa partie la plus immédiatement significative, généralement par des méthodes heuristiques, par suppression d'affixes ou de marques de flexion.

la même catégorie dans le thésaurus. Si c'est le cas, c'est un succès, dans le cas contraire, un échec. Les résultats sont en fait nuancés. Ils sont globalement corrélés avec les gammes de fréquences. Les contextes syntaxiques « écrémés », réduits aux relations de dépendance, donnent de meilleurs résultats pour les 600 mots les plus fréquents. Inversement, pour les formes moins fréquentes, les contextes pauvres débouchent sur davantage de succès. Cette variation tient en fait au nombre de traits disponibles dans chaque méthode pour partitionner les mots. Les contextes syntaxiques sont « maigres » et diminuent donc les éléments de rapprochement entre mots. Leur vertu ne va pas sans pauvreté... Seuls les mots très fréquents entrent dans suffisamment de contextes pour que cet élagage ne soit pas fatal. Par contre, les mots moins fréquents nécessitent des contextes plus larges pour disposer d'assez de points de convergence avec d'autres mots. L'expérience de Grefenstette conduit à penser qu'il n'est pas toujours nécessaire de recourir à une analyse syntaxique automatique pour obtenir des partitionnements satisfaisants. Le lien entre volume et nature des contextes est confirmé par Curran & Moens (2002). Cet article montre que du moment qu'on est en mesure d'augmenter significativement la taille du corpus en lui adjoignant des données similaires, les méthodes les plus simples (cooccurrences de lemmes ou de mots) deviennent aussi performantes que des méthodes plus complexes (analyse syntaxique).

4.2.2 Calcul du sens et perception sémantique

On peut également relativiser cette valorisation des contextes syntaxiques en questionnant son principe même. Elle repose en effet indirectement sur une conception calculatoire et compositionnelle du sens (à la Frege). Si le sens d'un groupe de mots est fonction de celui de ses composants (calculé à partir d'eux et des structures qui les organisent), *a contrario*, le groupe de mots pertinent (le syntagme) permet de cerner le sens d'un composant (en l'occurrence, l'ensemble des groupes de mots pertinents). Dans cette optique, un certain nombre de « ruses » viennent pallier les limites d'une définition pauvre du contexte. Réduire la fenêtre à quelques mots à gauche et à droite du mot à caractériser revient à fournir une version simpliste des contextes syntaxiques. Ainsi M. Rossignol écarte la caractérisation syntaxique des contextes et se contente de 1 à 3 mots à droite ou à gauche des pôles examinés. Le détail de la procédure (Rossignol, 2005 : 105), qui fait varier à la fois la taille de la fenêtre et les catégories qui y sont cherchées en fonction de la partie du discours dont relèvent les mots à classer revient tout de même un peu à faire entrer la syntaxe par la fenêtre (c'est le cas de le dire) après lui avoir montré la porte...

Dans le chapitre VIII de (Rastier, 1991), *La perception sémantique*, F. Rastier développe *l'hypothèse d'une unité fondamentale entre le perceptif et le sémantique* (p. 208). C'est l'intuition qu'au moins une partie des représentations sémantiques découlent moins d'un calcul que d'une reconnaissance de formes, c'est-à-dire de la mise en évidence de proximités entre des agglomérats de traits et des schémas d'ensemble qui forment des horizons d'attente. Ces schémas organisent la perception : c'est parce qu'il y a l'attente de tel ou tel schéma que tels ou tels traits sont perçus. C'est par exemple ce que montrent indirectement Valette & Grabar (2004) pour le repérage de contenus illicites ou préjudiciables sur le web (racisme en l'occurrence). Au sein d'un corpus de sites racistes et antiracistes, si les mots employés permettent de prédire le rattachement à l'une ou l'autre des catégories, d'autres traits, relevant d'autres niveaux, sont également contributifs, sinon discriminants : emploi des ponctuations, morphèmes sollicités, parties du discours privilégiées, etc. Le tableau 2 résume l'analyse.

Trait	*Ponctuation*	*Morphèmes*	*Parties du discours*
Sites racistes	! et …	*-ouille, -man, -phil*	Verbes
Sites antiracistes	; et ( )		Noms

Tableau 2 : Traits discriminants des sites racistes vs antiracistes

Des conclusions proches ont été tirées pour l'identification des grandes catégories de pages présentes sur le web (Beaudouin *et al*., 2002). Le repérage d'une isotopie relève aussi de ce travail indiciel : il y a attente d'isotopie(s) et repérage, en raison de cette attente, de convergences possibles de mots vers une ou des isotopie(s) possible(s). Sans doute peut-on faire une hypothèse complémentaire. Le travail indiciel, de reconnaissance de formes prend peut-être une place d'autant plus grande que l'unité sémantique et textuelle concernée s'élargit. Relativement restreint au niveau des contraintes de micro-syntaxe d'un mot, il prendrait une place prépondérante pour le repérage des thématiques et des genres. Cette hypothèse est en tout cas cohérente avec le relatif bon fonctionnement d'indices grossiers et relevant de niveaux multiples pour les thématiques, les genres et les styles (Karlgren & Cutting, 1994 ; Karlgren, 2000, Ivory & Hearst, 2002, Rossignol, 2005). En tout cas, ces hypothèses invitent à ne pas concevoir l'utilisation d'avatars (la représentation de contextes par de simples « mots ») comme une faiblesse, une pauvreté regrettable, mais au contraire comme une démarche cohérente avec la nature d'une partie des phénomènes sémantiques concernés. Elles invitent également à s'interroger sur les traits complémentaires (présentationnels, structurels par exemple) qui seraient constituables à partir des données utilisées et qui ne relèvent pas de l'analyse syntaxique ni de l'étiquetage morpho-syntaxique mais qui interviennent également dans la perception et la construction du sens.

5. Enseignements de l'analyse multilingue

Les analyses en corpus comparables ou parallèles sont présentées souvent – et nous n'y avons pas entièrement échappé dans la section 3 – comme une version dégradée de ce qui est possible en corpus monolingue. Nous souhaitons dans la présente section corriger partiellement cette représentation partiale, sinon désobligeante, c'est-à-dire montrer que les analyses en corpus monolingues peuvent tirer profit des enseignements issus du travail en corpus multilingues.

5.1. Rôle du lexique pivot

Le « lexique de transfert », ou « lexique pivot », employé dans l'alignement de corpus comparables, met en contact des mots des documents sources et cibles. Cette relation de traduction, spécifiée *a priori* par un lexique externe aux corpus, constitue une forme de supervision : on fournit à l'algorithme de recherche d'équivalents traductionnels une amorce pour son travail. Grâce à cette amorce, deux mots en relation de traduction, normalement considérés comme distincts, sont perçus comme équivalents par l'algorithme. Cela permet de « coller » l'un à l'autre les deux espaces lexicaux en présence, et ainsi de propager ces équivalences à d'autres couples de mots.

Le lexique pivot a un effet de « contraction » de l'espace des mots : deux mots jusque là différents se retrouvent considérés comme n'en faisant qu'un seul. Une même contraction pourrait être obtenue en monolingue par l'utilisation de dictionnaires de synonymes ou par

l'utilisation de classes de mots acquises. C'est ce que fait Schütze (1998) en recourant à la décomposition en valeurs propres (technique utilisée dans l'analyse sémantique latente ou LSA) pour diminuer l'espace des traits.

5.1.1 Constitution du lexique pivot

Fung & McKeown (1997) soulignent que tous les mots du lexique pivot n'ont pas les mêmes qualités pour servir de marqueurs de contexte. Elles réduisent donc ce lexique de transfert à un ensemble plus restreint de *mots amorces* (*seed words*) qui ont les propriétés suivantes : (*i*) une fréquence moyenne dans chaque corpus (entre 100 et 10 000 dans leur corpus anglais), pour avoir suffisamment de cooccurrences tout en évitant les mots qui cooccurrent avec trop de mots du corpus ; (*ii*) ne pas être un mot grammatical, de nouveau pour éviter les cooccurrences trop communes ; (*iii*) un faible taux de polysémie, un critère étant d'être traduction unique d'un mot de l'autre langue.

En effet, un mot dans une langue peut avoir plusieurs équivalents dans l'autre. Lorsqu'un mot du lexique pivot a ainsi plusieurs traductions, deux stratégies différentes sont observées : ne prendre en compte que l'une des traductions (Chiao & Zweigenbaum, 2002), ou toutes les prendre (Fung & McKeown, 1997 ; Déjean *et al.*, 2002). La première stratégie est plus simple ; la seconde, plus logique, implique de répartir les pondérations entre les différentes traductions. Une comparaison des deux stratégies reste à faire. Dans les deux cas, la place des termes complexes dans le lexique pivot pose problème. Après différentes expériences peu fructueuses de prise en compte de mots complexes, Déjean & Gaussier (2002) ont choisi de s'en tenir aux mots simples.

5.1.2 Utiliser les similarités avec le lexique pivot

Un inconvénient de la méthode « standard » de mise en correspondance de mots en corpus comparables est que si un mot n'a pour contexte aucun mot du lexique pivot, son vecteur de contextes est entièrement nul, et il n'est pas possible de lui trouver de traduction. Déjean & Gaussier (2002) proposent une méthode alternative fondée sur l'hypothèse suivante :

> *Deux mots de l_1 et l_2 sont, avec une forte probabilité, traduction l'un de l'autre si leurs similarités avec les entrées des ressources bilingues disponibles sont proches ».*

Le principe consiste à calculer des vecteurs de contextes pour les entrées du lexique pivot (ici un thésaurus bilingue médical, le MeSH) et à comparer le vecteur de contextes d'un mot source aux vecteurs de contextes des termes pivots. Cette comparaison calcule une similarité avec ces termes pivots, opérant une sorte de triangulation entre un mot source et les termes pivots distributionnellement les plus proches. Une triangulation homothétique dans l'espace cible devrait identifier les mots cible occupant une position dans la langue cible proche de celle du mot source dans la langue source.

Avec cette méthode, il n'est plus nécessaire de réduire le nombre de dimensions des vecteurs de contextes. Même si un mot du corpus ne cooccurre avec aucun des termes pivots, cela n'empêchera pas de comparer son vecteur de contextes aux vecteurs de contextes des termes pivots. Les expériences effectuées par Déjean & Gaussier (2002) rapportent des résultats un peu moins bons avec cette méthode qu'avec la méthode standard. En revanche, lorsqu'ils adaptent cette méthode pour prendre en considération les propriétés hiérarchiques du thésaurus MeSH, les résultats deviennent nettement meilleurs qu'avec la méthode standard. En effet, dans le thésaurus MeSH, les termes sont organisés dans une hiérarchie (où un même terme peut avoir plusieurs

pères)[20]. Le principe de cette adaptation est que lorsque deux termes pivots sont associés à un mot source, on propage cette association à leurs ancêtres communs et aux parents intermédiaires. Par exemple, si un mot est associé à *Hepatitis* et à *Cirrhosis*, on considérera qu'il est aussi lié à *Liver Diseases*.

Il faut souligner que cette méthode change le mode d'usage de la ressource pivot. On travaille non plus sur les cooccurrences d'un mot source avec les termes pivots, mais sur la proximité de son sens (à travers sa distribution, représentée par son vecteur de contextes) avec ceux des termes pivots (représentés eux aussi par leurs vecteurs de contextes). On peut alors se demander si cette méthode n'a pas un inconvénient inverse de la méthode standard. Si le sens d'un mot source est éloigné des sens de tous les termes pivots, n'ayant pas de points d'appui fiables sur lesquels effectuer la sorte de triangulation opérée, on peut s'attendre à obtenir des résultats moins pertinents. Le corpus sur lequel Déjean et Gaussier ont travaillé est constitué de résumés d'articles scientifiques médicaux tirés de la base Medline, et le thésaurus MeSH a été construit pour couvrir les besoins d'indexation de cette base documentaire. Il est donc possible que ce cas se produise peu. Les auteurs ne discutent pas ce point, mais proposent de combiner cette méthode avec la méthode standard (voir la section 5.2.3).

5.2. Croisements d'indices

Les deux méthodes de sélection d'équivalents traductionnels que nous venons de présenter s'appuient sur les distributions des mots dans les deux corpus. D'autres indices peuvent être employés pour aider cette mise en correspondance (sections 5.2.1 et 5.2.2). Il est aussi bénéfique de les combiner entre eux (section 5.2.3).

5.2.1 Filtrage a posteriori : similarité croisée

Le mode de comparaison de vecteurs de contextes mis en place est par nature asymétrique : si l'on part de la meilleure traduction candidate $(m_s)^c$ pour un mot source ms et que l'on cherche dans l'autre sens la meilleure traduction candidate $((m_s)^c)_s$, on ne retombe pas nécessairement sur le mot initial m_s. Observant cela, Sadat *et al.* (2003) et Chiao *et al.* (2004) combinent les informations obtenues en appliquant la recherche de traductions dans les deux directions, source → cible et cible → source. On peut faire un parallèle avec la recherche manuelle dans un dictionnaire bilingue : lorsque l'on a trouvé des traductions pour un mot source, il vaut mieux regarder ensuite dans la partie cible du dictionnaire les traductions proposées pour les mots cible obtenus : un mot cible qui a dans ses traductions le mot source initial a de meilleures chances d'en être une traduction plus centrale.

Sadat *et al.* calculent une nouvelle valeur de similarité entre mot source et mot cible en prenant le produit de leurs deux similarités directionnelles. Testé dans une tâche de recherche d'information translingue, ce reclassement leur fait gagner 27,1 % de précision moyenne (usage des cinq premières traductions en traduction de requête). Chiao *et al.*, de leur côté, travaillent directement sur le rang des traductions proposées : le nouveau rang d'une traduction est calculé

[20] Pour être plus précis, MeSH est organisé autour de « descripteurs ». Un descripteur est exprimé par un terme préférentiel et éventuellement des termes synonymes (« termes d'entrée »). Un descripteur peut être relié à des descripteurs plus larges (termes hyperonymes, holonymes, etc.) ou plus étroits (termes hyponymes, méronymes, etc.).

comme la moyenne harmonique[21] des deux rangs directionnels. Cela leur permet d'augmenter de 10 % le nombre de mots correctement traduits parmi les dix premières propositions.

5.2.2 Filtrage a posteriori : catégories morphosyntaxiques

Un mot d'une catégorie syntaxique donnée est souvent traduit par un mot de la même catégorie. C'est assez systématiquement le cas dans un dictionnaire bilingue. Cela peut l'être moins en corpus, où par exemple une construction *Nom de Nom* en français (*infarctus du myocarde*) pourra être traduite par une construction *Adjectif Nom* en anglais (*myocardial infarction*) : ici, *myocarde/Nom* est traduit en contexte par *myocardial/Adjectif* . Néanmoins, les possibilités de changement de catégorie restent limitées. Sadat *et al.* (2003) ajoutent donc un filtre de catégorie syntaxique sur les propositions de traduction anglais - japonais qu'ils obtiennent : Nom Nom, Verbe Verbe, et {Adjectif ou Adverbe}. Testé dans une tâche de recherche d'information translingue, ce filtrage supplémentaire leur fait gagner 11,5 % de précision moyenne.

Les expériences d'analyse distributionnelle monolingue menées par Bouaud *et al.* (1997) n'imposaient pas de contrainte sur les catégories syntaxiques des mots à comparer. Dans ces expériences, les vecteurs de contextes étaient construits à partir de relations de dépendance syntaxique. De ce fait, les regroupements de mots distributionnellement proches concernaient soit des noms, soit des adjectifs, mais pas les deux en même temps[22]. Avec une méthode employant non pas des dépendances syntaxiques, mais de simples sacs de mots, il est probable que noms et adjectifs pourraient se trouver mêlés. Un filtrage syntaxique *a posteriori* pourrait alors prendre son sens.

5.2.3 Combiner les ordres de similarité

Lorsque l'on dispose de plusieurs méthodes pour résoudre un problème, il est souvent plus productif de chercher à les combiner. C'est d'autant plus le cas lorsque les méthodes s'appuient sur des indices différents.

Déjean & Gaussier (2002) proposent de combiner avec la méthode standard leur méthode utilisant la similarité au lexique pivot (section 5.1.2). Les performances de la combinaison sont nettement supérieures à celles des méthodes individuelles, accroissant par exemple la f-mesure (moyenne harmonique du rappel et de la précision) de 50 à 84 % (méthode utilisant la hiérarchie ; un ensemble de propositions de traductions est considéré comme bon si une bonne traduction se trouve parmi les dix premières proposées). Dans un autre article sur ce travail, Déjean *et al.* (2002) indiquent qu'ils combinent également ces deux méthodes à l'usage direct du lexique pivot pour trouver des propositions de traduction. Dans ces travaux, ils considèrent que la probabilité de traduction d'un mot source par un mot cible est une combinaison linéaire des probabilités de traduction par chacun des modèles individuels. Les coefficients de cette combinaison linéaire sont appris sur une partie réservée du corpus, avec un ensemble distinct de mots dont la

[21] La moyenne harmonique est l'inverse de la moyenne des inverses : $\frac{1}{\frac{1}{2}\left(\frac{1}{x}+\frac{1}{y}\right)}=\frac{2xy}{x+y}$. Ainsi, si *foie* obtient *liver* comme deuxième traduction candidate et *liver* obtient *foie* en première position, le score croisé pour le couple {*foie, liver*} est $\frac{2\times2\times1}{2+1}=\frac{4}{3}=1,33$, ce qui est meilleur par exemple que celui obtenu pour le couple {*foie, lung*} (rang 1 dans une direction, 4 dans l'autre, d'où 1,6 au final). Cet indice favorise le fait d'être bien classé au moins dans l'une des deux directions.

[22] Ces mots se trouvaient dans des syntagmes nominaux, d'où l'absence ou la rareté de verbes et d'adverbes.

traduction est connue. Ici encore, les résultats combinés sont meilleurs que les résultats individuels.

En désambiguï sation sémantique automatique, les parties du discours des mots avoisinants renseignent sur le comportement syntaxique des mots en contexte, ce qui est une forme de combinaison d'indices. On voit mal par contre comment comparer les distributions de parties de discours d'une langue à l'autre, et donc comment s'en servir pour rapprocher les mots. En revanche, l'analyse distributionnelle pourrait être complétée par les relations repérées à l'aide de patrons lexico-syntaxiques : par exemple, les structures énumératives correspondent souvent à des relations de co-hyponymie.

6. Conclusion

6.1. Mode d'évaluation

Pour évaluer les traductions proposées par leur système, les auteurs constituent généralement un lexique bilingue servant de référence. L'évaluation se fait alors en comparant les traductions fournies par le système aux traductions de référence. Le lexique de référence peut être pris dans le lexique pivot (Chiao & Zweigenbaum, 2002) ou constitué manuellement à partir d'un extrait des mots du corpus (Déjean & Gaussier, 2002). Pour un mot source donné, les méthodes fournissent une liste ordonnée de traductions candidates. Comme il est difficile d'obtenir automatiquement une bonne traduction au premier rang de cette liste, il est habituel de s'intéresser aux mots source pour lesquels une traduction correcte se trouve parmi les n premiers mots cibles candidats : $n=10$ chez Déjean & Gaussier (2002), $n \in [1\ldots100]$ pour Fung & McKeown (1997), $n \in [1\ldots20]$ chez Chiao & Zweigenbaum (2002).

Le choix des mots de test est évidemment un paramètre crucial dans une telle évaluation. La tâche est plus facile avec des mots fréquents dans le corpus source (car leurs vecteurs de contextes seront mieux renseignés), et dont la traduction est fréquente dans le corpus cible ; elle est beaucoup plus difficile pour les mots moins fréquents, puisqu'ils auront à l'inverse peu de contextes. Si l'on dispose d'un lexique bilingue de bonne qualité, la méthode combinée incluant l'accès au lexique contribuera aux résultats de façon d'autant plus significative que les mots choisis seront inclus dans ce lexique : dans (Déjean & Gaussier, 2002), 48 % des mots obtiennent une traduction correcte par seul accès au lexique. Avec la méthode de similarité aux termes pivots (Déjean & Gaussier, 2002), les résultats devraient être meilleurs si les mots sont souvent des termes ou parties de termes du thésaurus pivot. Dans (Chiao & Zweigenbaum, 2002), qui utilisent la méthode standard, les tests se font successivement sur chaque mot du lexique pivot, que l'on supprime temporairement de ce lexique pour le test (« leave-one-out »).

Un autre type d'évaluation, centré sur la cohérence des résultats, est employé par Fung & McKeown (1997). Il consiste à construire des corpus comparables de même langue en divisant en deux un corpus initial (une collection d'articles du Wall Street Journal : WSJ 1993-1994). Dans ces conditions, on s'attend à ce qu'un mot du corpus source soit traduit par lui-même dans le corpus cible. Cela permet d'étalonner la méthode dans des conditions maîtrisées. Fung & McKeown montrent par exemple dans ces conditions que les résultats sont sensiblement meilleurs pour les mots moins polysémiques.

Notons que ces évaluations se font *in fine*, par rapport à une référence, sans intervention humaine au cours du traitement. On pourrait envisager également de recourir à un renforcement

positif (« relevance feedback ») par réintroduction des résultats jugés corrects dans le lexique pivot.

6.2. Quelle généralisation des méthodes ?

Les différents travaux effectués sur corpus comparables montrent qu'il est possible d'obtenir des propositions de traductions avec un niveau de qualité intéressant dans un certain nombre de situations. On peut néanmoins questionner la généralisabilité de ces résultats. Pour cela, examinons leurs conditions d'obtention.

Tout d'abord, le travail sur corpus comparables se présente en général par contraste avec le travail sur corpus parallèles. En pratique, on rencontre une gradation entre deux situations extrêmes (Déjean & Gaussier, 2002). À une extrémité, on peut utiliser un corpus parallèle sans tenir compte des informations d'alignement, ce qui donne un corpus comparable que l'on pourrait qualifier d'idéal. À l'autre extrémité, on placera des corpus comparables mais dont aucune paire de phrases n'est en relation de traduction, que l'on pourrait qualifier de corpus comparables au sens propre. On voit bien que dans le premier cas, même si l'on n'utilise pas les informations sur la correspondance entre phrases, le parallélisme entre les textes va donner de bonnes propriétés aux distributions des mots : les distributions de deux mots en relation de traduction auront de bien meilleures chances d'être très similaires que dans le cas de corpus comparables au sens propre. Les corpus employés par Déjean & Gaussier (2002) puis par Sadat *et al.* (2003) comportent une partie de textes parallèles : résumés Medline en anglais et en allemand des mêmes articles pour (Déjean & Gaussier, 2002), et 'abstracts' japonais et anglais de la collection de test NTCIR-2 pour (Sadat *et al.*, 2003). Ces travaux ont donc été réalisés dans des contextes particuliers – ce que reconnaissent leurs auteurs – et leur généralisabilité peut en dépendre. En revanche, le nombre de textes, voire de phrases parallèles dans les corpus comparables de Chiao & Zweigenbaum (2002), s'il y en a, est *a priori* très réduit, car il n'est pas constitutif de leurs couples de corpus. Et à l'extrême opposé, les corpus Wall Street Journal et Nikkei Financial News employés par Fung & McKeown (1997) sont, comme l'expriment les auteurs, «le type de corpus le plus non-parallèle », car ils ne partagent qu'un nombre limité de thèmes communs. Ce dernier exemple ajoute au non-parallélisme le handicap de différences de domaines et probablement aussi de genres discuté en section 4.1.

Le choix de la ressource pivot est lui aussi important. La qualité de la couverture du vocabulaire des corpus étudiés par le lexique ou la terminologie employée a certainement une influence sur les résultats obtenus. Comme nous l'avons indiqué en section 5.1.2, le thésaurus MeSH employé par Déjean & Gaussier (2002) (15 000 entrées anglais-allemand) est particulièrement approprié pour les résumés Medline qui constituaient leur corpus, ce qui est probablement une qualité importante pour la mise en correspondance par similarité avec les termes pivots. Chiao & Zweigenbaum (2003) étudient l'influence du lexique pivot employé. Partant d'un lexique bilingue comprenant essentiellement des mots du domaine médical (18 437 entrées français-anglais), ils lui ajoutent un lexique général (4 272 entrées). Les mises en correspondance sont meilleures avec le lexique combiné par rapport au lexique médical seul : cela montre qu'avec la méthode standard, les mots généraux contribuent aussi à décrire les contextes des mots du corpus, y compris ceux des mots médicaux. Ces expériences donnent à penser qu'un lexique possédant une bonne couverture du corpus, non seulement en mots du domaine mais également en mots 'généraux', est un atout pour la recherche d'équivalents traductionnels. Mais la dépendance des résultats à la composition de ce lexique (taille,

spécialisation, couverture par rapport au corpus traité), y compris selon les méthodes employées, reste à étudier plus précisément.

6.3. Nature des connaissances sémantiques acquises

(Mihalcea & Simard, 2005 : 239) soulignent la prise sur le sens qu'offrent les textes parallèles : *en l'absence d'alternatives en termes de « vraies » représentations sémantiques* (alternative 'true' semantic representations), *les textes parallèles nous offrent le moyen de découvrir le* sens *d'un texte, et de l'utiliser par voie de conséquence de différentes manières et pour des objectifs variés*. Les ressources de sémantique lexicale découlant des contextes multilingues, qu'il s'agisse de corpus comparables ou de corpus alignés, ne sont pas des traductions dans un formalisme quelconque, mais des mises en relation, souvent bruitées, de mots et de séquences. Elles proposent en fait des paraphrases possibles, qu'il reste à trier et valider. On retrouve la conception défendue par I. Mel'cuk des paraphrases comme les meilleures représentations possibles du sens (Mel'èuk, 1998). Peut-être vaut-elle également pour le travail fait en contexte monolingue...

Références

BEAUDOUIN V., FLEURY S., HABERT B., PASQUIER M. & LICOPPE C. (2002) « Décrire la Toile pour mieux comprendre les parcours. Sites personnels et sites marchands »,dans Valérie Beaudouin, Christian Licoppe (ed.), *Parcours sur Internet*, revue *Réseaux*, 20 (116), p.19-51.

BIBER D. (1988) *Variation accross speech and writing*, Cambridge : Cambridge University Press.

BIBER D. (1993) « Using register-diversified corpora for general language studies », dans *Computational Linguistics*, 19(2), p.243-258.

BIBER D. (1995). Dimensions of register variation : a cross-linguistic comparison. Cambridge : Cambridge University Press.

BOUAUD J., HABERT B., NAZARENKO A. & ZWEIGENBAUM P. (1997), « Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles », dans *Actes des 1ières journées Ingénierie des Connaissances*, p.207-223, Roscoff, France.

BOURIGAULT D. (2002) *UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus*, dans *Actes de TALN'02*, p. 75-84, Nancy : ATALA.

CHIAO Y.-C., STA J.-D. & ZWEIGENBAUM P. (2004) « A novel approach to improve word translations extraction from non-parallel, comparable corpora », dans *Actes International Joint Conference on Natural Language Processing*, Hainan, China : AFNLP.

CHIAO Y.-C. & ZWEIGENBAUM P. (2002) « Looking for candidate translational equivalents in specialized, comparable corpora », dans *Actes 19th COLING*, p. 1208-1212, Taipei, Taiwan.

CHIAO Y.-C. & ZWEIGENBAUM P. (2003) « The effect of a general lexicon in corpus-based identification of French-English medical word translations », dans R. BAUD, M. FIESCHI, P. LE BEUX & P. RUCH (eds.), *Actes Medical Informatics Europe*, volume 95 of Studies in Health Technology and Informatics, p.397-402, Amsterdam : IOS Press.

CURRAN J. R. & MOENS M. (2002) « Scaling context space », dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL'02), p. 231–238, Philadelphia.

DAGAN I., ITAI A. & SCHWALL U. (1991) « Two languages are more informative than one », dans *Actes ACL 1991*, p.130-137 : ACL.

DARMONI S. J., LEROY J.-P., THIRION B., BAUDIC F., DOUYERE M. & PIOT J. (2000. « CISMeF : a structured health resource guide », dans *Methods of Information in Medicine*, 39(1), p.30-35.

DÉJEAN H. & GAUSSIER E. (2002) « Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables », dans VERONIS J. (resp) *Lexicometrica*. *Numéro spécial Alignement lexical dans les corpus multilingues*.

DÉJEAN H., GAUSSIER E. & SADAT F. (2002) « An approach based on multilingual thesauri and model combination for bilingual lexicon extraction », dans *Actes 19th COLING*, Taipei, Taiwan.

FERRET O. (2004) « Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales », dans B. BEL & I. MARLIEN (eds.), *TALN 2004 XIe conférence sur le traitement automatique des langues naturelles*, Fès (Maroc) : ATALA (Association pour le Traitement Automatique des Langues).

FIRTH J. (1957) « A synopsis of linguistic theory 1930-1955 », dans *Studies in Linguistic Analysis*, p.82-95. Réédité, *Selected Papers of J. R. Firth*, F. Palmer (ed), Longman.

FUNG P. & MCKEOWN K. (1997) « Finding terminology translations from parallel corpora », dans *Actes Fifth Annual Workshop on Very Large Corpora*, p.192-202 : ACL.

GREFENSTETTE G. (1994a) « Corpus-derived first, second and third order affinities », dans *EURALEX*, Amsterdam.

GREFENSTETTE G. (1994b) *Explorations in Automatic Thesaurus Discovery*, Dordrecht, The Netherlands : Kluwer Academic Publisher.

GREFENSTETTE G. (1996) « Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches », dans B. BOGURAEV & J. PUSTEJOVSKY (eds.), *Corpus Processing for Lexical Acquisition, Language, Speech and Communication*, chapitre 11, p.205-216. Cambridge, Massachusetts : The MIT Press.

GROSS G. (1994) « Classes d'objets et description des verbes », dans *Langages* (115), p.15-30.

HABERT B. (2005) *Instruments et ressources électroniques pour le français*. Collection L'essentiel français. Gap/Paris : Ophrys.

HABERT B., GRABAR N., JACQUEMART P. & ZWEIGENBAUM P. (2001) « Building a text corpus for representing the variety of medical language », dans *Actes Corpus Linguistics*, Lancaster : UCREL.

HABERT B. & ZWEIGENBAUM P. (2002) « Régler les règles », dans TAL, 43(3), p.83-105. Problèmes épistémologiques. M. Cori, S. David, J. Léon (resp.).

HARRIS Z. S. (1991) *A theory of language and information. A mathematical approach*. Oxford : Oxford University Press.

HERSH W. R., BALL A., DAY B., MASTERSON M., ZHANG L. & SACHEREK L. (1999). « Maintaining a catalog of manually-indexed, clinically-oriented World Wide Web content », dans *Journal of the American Medical Informatics Association*, 6 (suppl), p.790-794.

ILLOUZ G., HABERT B., FLEURY S., FOLCH H., HEIDEN S. & LAFON P. (1999). « Maîtriser les déluges de données hétérogènes », dans A. CONDAMINES, C. FABRE &

M.-P. PÉRY-WOODLEY (éds.), *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, p.37-46, Cargèse.

IVORY M. & HEARST M. (2002), « Statistical profiles of highly-rated web sites », dans *CHI 2002, ACM Conference on Human Factors in Computing Systems.*

JARDINO M. (2004) « Recherche de structures latentes dans des partitions de « textes » de 2 à *k* classes », dans G. PURNELLE, C. FAIRON & A. DISTER (éds.), *Le poids des mots. Actes des 7èmes journées internationnales d'analyse statistique des données textuelles*, volume 2, p.661-671, Louvain-la-Neuve, Belgique : UCL Presses universitaires de Louvain.

KARLGREN J. (1999). Stylistic experiments in information retrieval. In T. STRZALKOWSKI, éditeur, Natural language information retrieval, Text, speech and language technology, chapitre 6, p. 147–166. Dordrecht : Kluwer.

KARLGREN J. (2000). Stylistic Experiments for Information Retrieval. Phd in computational linguistics, Sweedish Institute of Computer Science, Stockolm, Sweden.

KARLGREN J. & CUTTING D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In Proceedings 15th International Conference on Computational Linguistics (COLING'94), Kyoto.

KILGARRIFF A. & GREFENSTETTE G. (2003) «Introduction to the special issue on the Web as a corpus », dans *Computational Linguistics*, 29(3), p.333-347.

LE PESANT D. (1994) « Les compléments nominaux du verbe lire : une illustration de la notion de 'classe d'objets' », dans *Langages* (115), p.31-46.

LEBART L., MORINEAU A. & PIRON M. (1997) *Statistique exploratoire multidimensionnelle*. 2e cycle, Paris : Dunod, 2ème edition.

LIN D. & PANTEL P. (2002) *Concept discovery from text,* dans *COLING'02*, p.577-583, Taipei, Taiwan.

LOSEE R. M. (1998) *Text Retrieval and Filtering : Analytic Models of Performance. Information Retrieval*. Dordrecht : Kluwer Academic Publishers.

MANNING C. D. & SCHÜTZE H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts : The MIT Press.

MEL'CUK I. (1988) « Paraphrase et lexique dans la théorie linguistique sens-texte », dans *Lexique* (6), p.13-54.

MIHALCEA R. & SIMARD M. (2005) «Parallel texts », dans *Natural Language Engineering* 11(3), p.239-246.

PICHON R. & SÉBILLOT P. (1999) « Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience », dans P. AMSILI (éd.), *Actes TALN'99*, p.279-288, Cargèse : ATALA.

RAPP R. (1995) « Indentifying word translation in non-parallel texts », dans *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, student session, volume 1, p.321-322, Boston, Mass.

RAPP R. (1999) « Automatic identification of word translations from unrelated English and German corpora », dans Actes 37th ACL, College Park, Maryland.

RASTIER F. (1987) *Sémantique Interprétative.* Paris : PUF.

RASTIER F. (1991) *Sémantique et recherches cognitives. Formes sémiotiques*. Paris : Presses Universitaires de France.

RASTIER F., CAVAZZA M. & ABEILLé A. (1994) *Sémantique pour l'analyse : de la linguistique à l'informatique*. Sciences Cognitives. Paris : Masson.

RESNIK P. & SMITH N. A. (2003) « The Web as a parallel corpus », dans *Computational Linguistics*, 29(3), p.349-380.

ROSSIGNOL M. (2005) *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. Thèse de doctorat en informatique, Université de Rennes 1, Rennes.

SADAT F., YOSHIKAWA M. & UEMURA S. (2003) « Learning bilingual translations from comparable corpora to cross-language information retrieval : Hybrid statistics-based and linguistics-based approach », dans J. ADACHI & K.-F. WONG (éds.), *Actes Sixth International Workshop on Information Retrieval with Asian Languages*, p.57-64.

SCHÜTZE H. (1998). « Automatic word sense discrimination », dans *Computational Linguistics*, 24(1), p.97-124.

VALETTE M. & GRABAR N. (2004) « Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP », dans G. PURNELLE, C. FAIRON & A. DISTER (éds.), *Le poids des mots. Actes des 7èmes journées internationales d'analyse statistique des données textuelles*, volume 2, p.1106-1116, Louvain-la-Neuve, Belgique : UCL Presses universitaires de Louvain.

VÉRONIS J. (2000a) « Alignement de corpus multilingues », dans J.-M. PIERREL (éd.) *Ingénierie des langues, Informatique et systèmes d'information*, chapitre 6, p.151-172. Paris : Hermès Science.

VERONIS J. (éd.) (2000b) *Parallel Text Processing : Alignment and use of translation corpora.* Dordrecht : Kluwer Academic Publishers.

VERONIS J. (2004) « HyperLex : Lexical cartography for information retrieval », dans *Computer Speech and Language*, 18(3), p.223-252.

SYSTEME DE RECHERCHE D'INFORMATION MEDICALE PAR CROISEMENT DE LANGUES : VIETNAMIEN-FRANÇAIS-ANGLAIS[23]

Tuan Duc Tran,

DYALANG, Université de Rouen

Introduction

La performance d'un système de recherche d'information médicale par croisement de langues réside dans sa capacité de produire à la suite d'une requête en sa langue maternelle des documents aussi pertinents en langue source qu'en langues cibles dans lesquels le terme requis apparaît et circule en fonction de sa hiérarchie notionnelle et domaniale. Dans la recherche des informations médicales, les utilisateurs vietnamiens sont souvent confrontés à deux niveaux de difficultés. Le premier résulte de l'explosion des variantes terminologiques produites soit par la vulgarisation scientifique soit par la pratique terminologique des spécialistes. La deuxième difficulté réside dans la pénurie des documents en vietnamien et des ressources terminologiques médicales en ligne. Les laboratoires de recherche sur le traitement automatique de la langue vietnamienne se heurtent à des difficultés d'acquisition de corpus spécialisés et d'outils pour acquérir automatiquement des termes à partir de textes. L'extraction des termes vietnamiens reste encore en phase d'expérimentation. Dans l'objectif d'intégrer une modélisation des concepts médicaux en vietnamien inspirée de la socioterminologie dans le système de recherche d'information médicale par croisement de langues qui tient compte du processus de l'acculturation des termes, de la variation terminologique, nous proposons de construire un système de recherche d'information médicale par croisement de langue dans lequel l'utilisateur vietnamien peut effectuer des recherches documentaires en sa langue maternelle et obtenir des documents pertinents en français ou en anglais via un module dictionnairique trilingue. Cet article s'articule en 3 sections. Nous étudions dans la section 1 l'approche socioterminologique qui sous-tend l'observation du terme circulant dans son contexte, du microcontexte au texte intégral. Nous examinons dans la section 2 les particularités de la langue vietnamienne ainsi que les modèles de la formation terminologique médicale. Nous décrivons dans la section 3 la

[23] L'article présenté n'a pas pu encore réaliser les validations en TAL des données du vietnamien, le travail présenté ici est donc un travail en cours.

composition des modules du système de recherche d'information médicale (SRIM) par croisement de langues : vietnamien >français>anglais.

1. Méthode et constitution du corpus

1.1. Pourquoi une jonction entre socioterminologie et SRIM ?

La socioterminologie est une terminologie d'orientation sociolinguistique qui « procède avant tout d'une attitude descriptive » (Gaudin, 1992, p.156). Cette nouvelle discipline prend en compte les aspects sociolinguistiques de la communication scientifique et technique. Elle s'intéresse aux « pratiques institutionnelles qui visent l'observation, l'enregistrement et la normalisation des pratiques langagières dans les procès technologiques » (Dubois et al., 1994, p.436). Du point de vue épistémologique, la socioterminologie insiste sur « les pratiques langagières, et non plus sur la seule «langue » réglée des experts et des normes » (Gaudin, 1993, p.247) et prend en compte la description des usages réels des termes.

1.2. Approche socioterminologique

En tenant compte du fonctionnement réel du terme, la *socioterminologie* est une approche terminologique qui s'est penchée sur la circulation sociale des termes dans leurs contextes de spécialisation. Il s'agit d'une terminologie réconciliée avec l'usage, elle constate que «*faire de la terminologie suppose de s'interroger à la fois sur l'aspect conceptuel et sur l'aspect discursif* » (Bouveret, 1996, p.48). L'approche socioterminologique est une approche sociolinguistique en terminologie qui relie la pratique glottopolitique avec les usages sociaux. La variation terminologique est un phénomène très répandu en terminologie médicale vietnamienne comme les exemples cités ci-dessus, nous jugeons souhaitable de recourir à la socioterminologie pour tenter d'étudier le fonctionnement des termes médicaux vietnamiens considérés comme unités textuelles et non plus comme des termes indépendants du contexte discursif. Le terme est à la fois un signe linguistique et une représentation de connaissance, un concept qui se définit explicitement en référence à un domaine de savoir (Lerat, 1995, p.5). L'interprétation du contenu de connaissance d'un terme ne peut pas être effectuée hors du contexte et un terme doit donc être observable dans le corpus. La socioterminologie a pour objectif « *de se soucier du fonctionnement des termes et des conditions sociolinguistiques* » (Gaudin, 1990, p.14). En effet, fondée sur une approche sociolinguistique, la socioterminologie met l'accent sur « *les pratiques langagières et non plus sur la seule langue réglée par des experts et des normes ; refus de l'amalgame entre sciences et techniques au profit d'une approche plus fine et contrastive ; primat accordé à la description sur la prescription dans l'intervention des linguistes ; prise en compte de la dimension industrielle de la communication "scientifique et technique"* » (Gaudin, 1992, p.152). La terminologie médicale vietnamienne a une forte tendance descriptive qui facilite le processus d'acculturation du terme. En effet, l'attitude descriptive est « *une attitude plus linguistique* » qui suppose que les termes médicaux soient étudiés *dans leur dimension discursive* (ibid.). L'objectif de la socioterminologie est de chercher à réintroduire la terminologie dans la pratique socio-discursive et à comprendre le lien entre la dimension sociale des terminologies prises dans les relations de concurrence, de pouvoir et la dimension linguistique-cognitive (Gambier, 1993, p.103). Si l'on considère l'acculturation du terme comme le mécanisme de socialisation, un processus d'intégration et de démocratisation du savoir à travers le système

d'information informatisé, le système de recherche d'information médicale par croisement de langues devient un outil de validation « socialisé » de la pertinence du terme requis.

1.3. Constitution du corpus

Pour le corpus, nous avons constitué un corpus comparable à partir de textes médicaux (Tran, 2003). En matière de recherche d'information médicale, ce projet dictionnairique répondra à la fois aux besoins des utilisateurs qui font une requête en vietnamien et obtiennent des documents vietnamien, français, anglais et à la construction des ressources terminologiques en vietnamien dans le but de transfert de connaissances, de vulgarisation de connaissances biomédicales.

Notre réflexion s'appuie sur une application concrète de construction d'un dictionnaire médical trilingue dans le cadre de la recherche d'information et de la traduction médicale. Notre corpus comparable est constitué d'un ensemble de trois corpus monolingues (vietnamien/ français/ anglais). Selon Déjean et Gaussier (2002), «*l'utilisation de corpus comparable permet d'avoir accès directement à la terminologie monolingue originelle d'un domaine, à l'usage réel des mots dans chaque langue, et évite donc le biais introduit par la traduction* ». Les textes qui vont composer le corpus doivent répondre à deux critères de pertinence : vis-à-vis du domaine médical (cours ou articles de référence dans la littérature médicale vietnamienne) ; vis-à-vis de l'application (recherche d'information et traduction médicale). Nous faisons l'hypothèse que le corpus comparable est composé d'une quantité suffisante de textes pour nous permettre de construire un dictionnaire médical au service de la recherche d'information médicale et de la traduction médicale. Dans un corpus comparable, le contexte est utilisé comme un outil de validation de la plausibilité d'une traduction. En terminologie descriptive, le contexte nous donne un accès au contenu de connaissance du terme ; en particulier il nous permet d'éviter certaines erreurs d'interprétation et nous fournit une description compréhensible du fonctionnement du terme. De plus, on voit que le sens est un processus de déroulement constant qui se construit tout au long du discours. Le contexte permet de désigner le sens d'un terme et peut servir de cadre de désambiguï sation sémantique du terme donné. D'après notre expérience, un lecteur qui n'est pas spécialiste du domaine est souvent intéressé par des relations lexicales qui lui permettent de structurer les connaissances du domaine, de les catégoriser pour construire le sens du terme.

Le niveau des textes répond à la nature et à la destination du savoir transmis. Il existe en effet plusieurs niveaux de textes correspondant à la transmission soit d'un savoir constitué soit d'un savoir en cours de constitution. Autrement dit, le premier type de texte est didactique, destiné à la formation des médecins généralistes, les textes sur lesquels nous nous sommes basé correspondent aux connaissances médicales d'un médecin généraliste telles que la physiologie médicale, la génétique relatives aux sciences fondamentales et les maladies infectieuses liées aux maladies bactériennes, virales et parasitaires. Le second type de texte est cognitif ; il a pour but de présenter des connaissances scientifiques en cours d'actualisation. Il s'agit d'une part, des articles de recherche publiés de 1998 à 2005 dans les revues de référence (version électronique en ligne) : *T p chí y d c h c*[24] (Revue médico-pharmaceutiques), *T p chí y h c th c hành*, *T p chí D c h c*, *T p chí thông tin y h c*[25], publiées par Institution Centrale de l'information et de la bibliothèque médicale, *T p chí y h c Thành phô Hô Chi Minh*[26] ; d'autre part, des cours de médecine. Le corpus vietnamien est composé d'environ 1600 articles, le corpus français plus de

[24] http://www.ykhoa.net/tapchiyhoc/index.htm
[25] http://www.cimsi.org.vn/
[26] http://tcyh.yds.edu.vn/

700 documents en lignes de type d'accès libre (CISMeF[27], UMVF[28]), le corpus anglais de PubMed Central (PMC[29]).

2. Particularités de la langue vietnamienne

Pour comprendre mieux la formation terminologique en vietnamien, nous aborderons dans cette partie des particularités de la langue vietnamienne liées principalement à la création lexicale vietnamienne. Appartenant à la famille de langues austroasiatiques (Haudricourt, 1953, p.122-128), la langue vietnamienne est l'un des spécimens les plus typiques des langues syllabiques et isolantes : la plupart du temps, le mot, le morphème et la syllabe coïncident presque entièrement (Cao, 1985, p.17). La langue vietnamienne présente trois caractéristiques principales : c'est une langue à tons, une langue isolante, non flexionnelle, et une langue à classificateurs. Les mots peuvent se juxtaposer sans mots-outils. La juxtaposition est un des modèles de construction syntaxique importants. La spécification notionnelle en vietnamien est affaire de syntaxe par exemple : *b nh nhân cao huyêt 'ap* (hypertendu). La plupart des linguistes vietnamologues admettent que le morphème est l'élément de base constituant les mots vietnamiens (Hoàng et al., 1998, p.23). Pourtant le vietnamien n'est pas vraiment une langue monosyllabique. C'est la raison pour laquelle dans le traitement automatique du langage naturel, la segmentation en mots à partir d'un corpus textuel devient difficile en particulier dans la désambiguïsation des syntagmes nominaux extraits. Le sens du syntagme nominal extrait dépend fortement du contexte donné. Dans la perspective de construction d'un lexique médical vietnamien, cette analyse consiste à distinguer deux sortes de mots. Une première distinction sépare les mots immotivés, c'est-à-dire inanalysables, constitués d'un seul morphème ou monosyllabe (ex. *b nh/maladie*), de ceux qui sont relativement motivés, c'est-à-dire analysables, constitués de deux morphèmes ou deux syllabes pour former un mot composé dissyllabique (ex. *b nh cãn/étiologie*). En général, dans le fonds lexical vietnamien, on peut distinguer trois types de composés populaires : les composés à sens synergique, les composés à sens analytique, et les composés par redoublement. Ces composés apportent souvent une spécificité, en sorte qu'il n'est pas possible de tenir leurs sens pour la seule somme des sens des composants. En plus, on estime à 70% en moyenne la présence de vocables sino-vietnamiens ou *hán-vi t* dans le fonds lexical vietnamien (Li & Waters, 1998, p.IX). Ces ressources de vocabulaire contribuent activement à la formation des mots composés savants (ex. *y h c/médecine*) et des syntagmes nominaux (ex. *thông tin y h c /information médicale*). Les composés savants représentent approximativement 22% de la terminologie médicale vietnamienne (Tran, 1999) et les termes syntagmatiques 75% (Tran, 2002).

3. Vers un système de recherche d'information médicale par croisement de langues basé sur un dictionnaire médical vietnamien >français>anglais

3.1. Extraction des termes et construction dictionnairique de base

Certains spécialistes vietnamiens en linguistique computationnelle constatent que l'extraction automatique des termes vietnamiens axée sur des bases statistiques n'a pas encore atteint le

[27] http://www.chu-rouen.fr/cismef/
[28] http://www.med.univ-rennes1.fr/UVMF/UMVF3/
[29] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PMC

niveau de développement avancé (Le A. Ha, 2003). Dans la condition où l'extracteur n'est pas disponible, les termes médicaux vietnamiens sont manuellement extraits par un linguiste spécialisé en langage médical après avoir effectué un pré-traitement de texte avec l'outil *Unitex*[30] (corpus processing system) de l'Université de Marne-la-vallée. Cet outil permet de segmenter le corpus textuel en phrases**.** Le terme qui apparaît dans un contexte définitoire est extrait de préférence grâce à sa richesse en information sur les relations lexicales qui donnent accès au modèle conceptuel. Dans la formation des termes médicaux vietnamiens, le mode de formation par syntagme représente 61,24% des termes extraits. Le syntagme nominal est une préoccupation prioritaire dans la technologie langagière vietnamienne. A cet égard, le syntagme terminologique reste un remarquable outil de dénomination. Dans sa thèse sur une indexation structurée basée sur des syntagmes nominaux, Ho B. Q (2004) a construit un analyseur vietnamien pour catégoriser les mots et pour extraire des syntagmes nominaux. Cet auteur a proposé une approche de recherche d'information bilingue basée sur des syntagmes nominaux avec une liaison entre des mots-tête mais son travail reste encore en phase d'expérimentation. Nous avons recensé une liste de 30 000 termes-candidats vietnamiens.

Pour les corpus français et anglais, nous avons utilisé les modules *ExtractTerm* et *OrganizerTerm* de *Xerox Terminology Suite*[31] (XTS) version 2 (Xerox Research Centre Europe en France) pour extraire automatiquement les candidats-termes (Tran et al,2003a). L'outil XTS est un outil de construction automatique de terminologie réalisant une analyse morpho-syntaxique des phrases pour en extraire les syntagmes nominaux (fig.1). Après avoir extrait les termes-candidats à partir de corpus comparable, le travail de l'appariement des termes est un travail manuel. Le dictionnaire est constitué d'un ensemble de fiches, chaque fiche peut être modifiée par une ou plusieurs personnes dûment mandatées. L'outil de gestion du lexique est une base de données *Access 2000*. Au total, la base pourra contenir 30 000 entrées (Tran et al., 2004a). Les relations sémantiques nous amènent à reconnaître le champ conceptuel du terme et permettent de différencier des termes équivoques dans le même système.

[30] http://igm.univ-mlv.fr/~unitex/
[31] http://www.mkms.xerox.com/

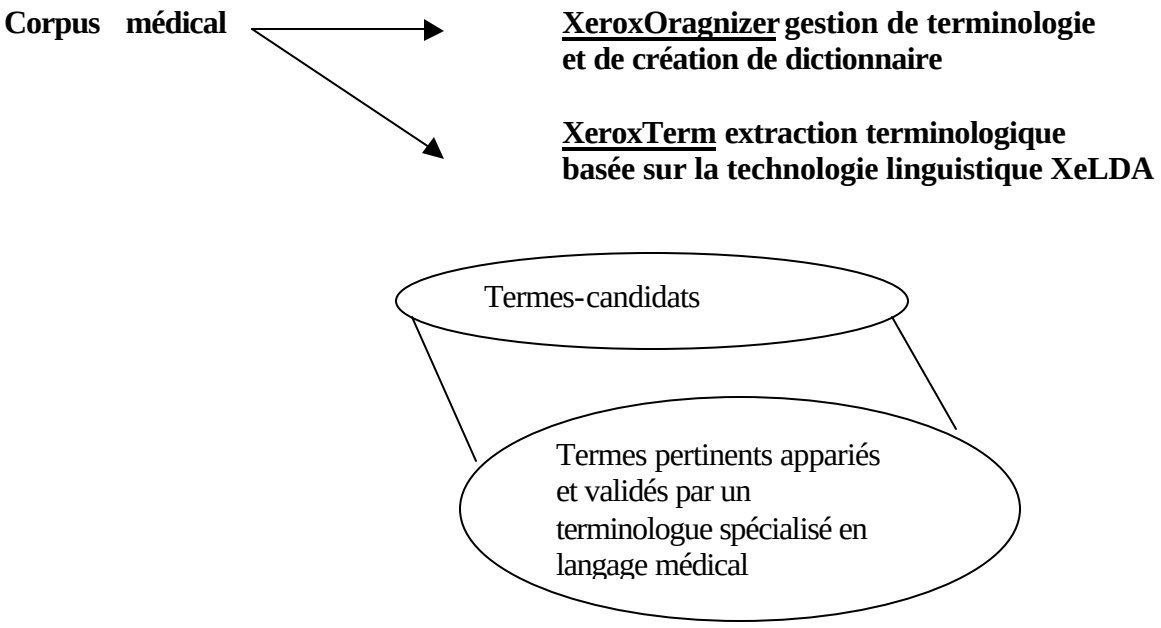

Figure 1. Schéma d'extraction terminologique

3.2. Système de recherche d'information médicale par croisement de langues

La technique de recherche d'information par croisement de langues s'applique aux moteurs de recherche multilingues : on formule une requête dans sa langue maternelle (vietnamien par exemple) pour récupérer des documents écrits dans d'autres langues différentes de celle de la requête (français ou/et anglais par exemple). Actuellement, les approches adoptées se basent principalement sur des processus de traduction de requêtes (query translation) ou de traduction de documents (document translation) ; le but étant de représenter les documents et/ou les requêtes dans un même référentiel. Cette traduction peut être obtenue en utilisant des traducteurs automatiques, un vocabulaire spécifique ou des dictionnaires (Tran et al, 2004b). Les travaux de recherche dans ce domaine se sont principalement orientés sur la traduction de requêtes. Notre travail se focalise sur la traduction des termes de requête en utilisant notre lexique trilingue comme moyen de traduction (Tran et al, 2004a). Un formulaire permet à l'utilisateur de saisir un mot ou une expression en vietnamien, français ou anglais. Une fois que l'utilisateur valide le formulaire, l'outil Internet cherche une correspondance dans le lexique et lance une recherche sur les agents de recherche d'information médicale de référence (CISMeF/Doccismef, PubMed) pour mot clé avec la traduction. La requête traduite va chercher un appariement avec l'UMLS[32]. Le résultat sera la traduction du mot saisi dans le langage choisi et la liste des sites obtenus à partir de l'agent de recherche d'information choisi (Fig. 2).

[32]Metathesaurus Unified Medical Language System : http://www.nlm.nih.gov/research/umls/

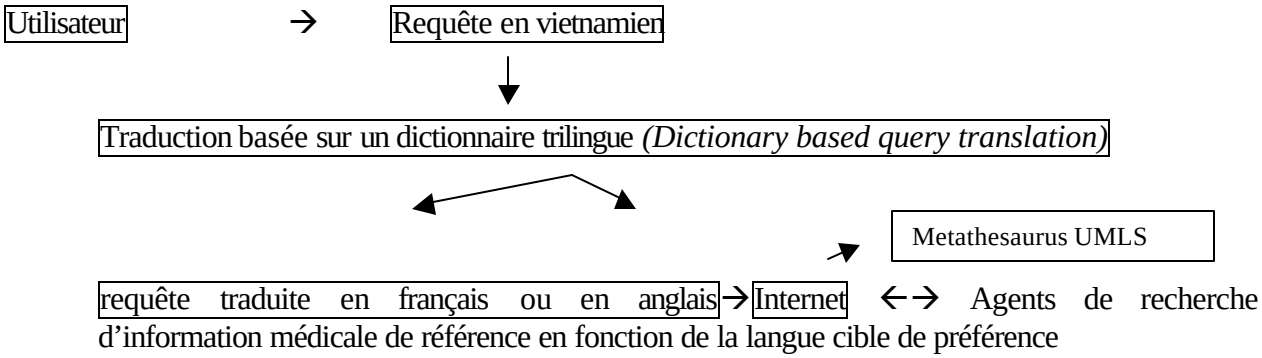

Fig.2 : Structure du système de recherche d'information médicale par croisement de langues

Conclusion

Dans l'optique de la socioterminologie, une approche sociolinguistique de la terminologie, nous nous sommes attaché à la description sémantique qui favorise une modélisation des concepts médicaux en vietnamien. Dans les échanges médico-pharmaceutiques au niveau international ou multinational, le SRIM par croisement de langues couplé avec le système de recherche documentaire de l'Université Médicale Virtuelle Francophone (UMVF)[33] (Tran et al., 2003b) et avec accède au Doc'CISMeF[34], sites et documents recensés dans CISMeF montre d'une part, la dynamique sociale de la francophonie et d'autre part, est connecté avec des ressources bibliographiques en anglais par le biais de PubMed. Le SRIM par croisement de langues est un système qui vise dès l'origine à répondre aux besoins de recherche d'informations médicales et de traduction des professionnels de santé. Il a pour but enfin de proposer un modèle variationniste de l'implantation terminologique au Vietnam.

Références

Bouveret M., 1996, *Néologie et terminologie : Production de sens du terme*, Thèse de Doctorat (Tome I) dir. Paul Siblot, Université Paul Valéry - Montpellier III.

Cao Xuân Hao, 1985, « Les linguistes vietnamiens et la phonologie de leur langue », dans *Phonologie et linéarité*, SELAF, n° spécial 18, Paris, pp. 260-277.

Déjean H, Gaussier E., 2002, « Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables », dans *Lexicometrica*, No spécial 2002, pp. 22 (version électronique).

Dubois J. et al., 1994, *Dictionnaire de linguistique et des sciences du langage*, ed. Larousse.

Gambier Y., 1993, « Implications épistémologiques et méthodologiques de la socioterminologie », dans *Actes du XVe Congrès International des Linguistes*, Québec, Université de Laval, 9-14 Août 1993, vol. 2, Presses de l'Université de Laval, pp. 99-113

Gaudin F., 1990, *Terminologie : des problèmes sémantiques aux pratiques institutionnelles*, Thèse de doctorat, 2 vol., dir. Louis Guespin, Université de Rouen.

[33] http://www.med.univ-rennes1.fr/UVMF/UMVF3/index.php
[34] http://doccismef.chu-rouen.fr/

Gaudin F., 1992, « L'apparition de la socioterminologie : une position épistémologique » dans *Où en sont les sciences du langage 10 ans après ?*, ed. ASL, pp 151-160

Gaudin F., 1993, *Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles*, Publication de l'Université de Rouen.

Gaudin F., 2003, *Socioterminologie : une approche sociolinguistique de la terminologie*, Champs linguistiques/ Manuels, De boeck Duculot, Bruxelles.

Guespin L. et Laroussi F., 1989, « Glottopolitique et standardisation terminologique », dans *La banque des mots*, numéro spécial, ed. CILF, pp. 5-21.

Haudricourt A.G., 1953., « La place du vietnamien dans les langues austroasiatiques », Bulletin de la Société de Linguistique de Paris, 49, 1

Hoàng V n Hành et al., 1998, *T ti ng Vi t : Hình thái, c u trúc, t láy,t ghép, chuy n lo i* (Les mots vietnamiens : morphologie, structure, mots redoublés, mots composés, transition), Centre National des Sciences sociales et humaines, Institut de Linguistique, Maison d'édition-Sciences sociales.

Ho BQ, 2004, *Vers une indexation structurée basée sur des syntagmes nominaux (impact sur un SRI en vietnamien et la RI multilingue)*, Thèse Doctorat Université Joseph Fourier-Grenoble I.

Le An Ha, 2003, "A method for word segmentation in Vietnamese", dans *Proceedings of Corpus Linguistics 2003*, Lancaster, UK.

Le Beux P et al., 2001, UMVF, «Définition des spécifications du projet de l'Université Médicale Virtuelle Française », dans *AIM 2001 : Télémédecine et eSanté.*

Lehmann A. et Martin-Berthet F., 1998, *Introduction à la lexicologie: Sémantique et morphologique*, Dunod, Paris.

Lerat P., 1995, *Les langues spécialisées*, P.U.F, Paris.

LI Leyi (LY Lac Nghi) et Jim Waters, 1998, *In search of the origins of Chinese characters relevant to Vietnamese*, Nxb Thê Gioi, Hanoi.

Tran D.T. 1999, *Standardisation de la terminologie médicale vietnamienne : une approche socioterminologique*, Thèse de doctorat, dir. B. Gardin et F. Gaudin, Université de Rouen.

Tran D.T., 2002, « Typologie des constructions syntagmatiques des termes médicaux vietnamiens : une tentative de la démocratisation du savoir », dans *Terminology, International Journal of Theoretical and Applied Issues in Specialized Communication* vol. 8, numéro 2, John Benjamins publishing Company, pp.207-220.

Tran D.T et al., 2003a, « Acquisition semi-automatique de terminologie bilingue en biologie moléculaire à partir des corpus comparables », dans *Actes des Ve rencontres TIA 2003*, LIIA-ENSAIS, Strasbourg. pp.166-175.

Tran DT et al, 2003b, « Indexation semi-automatique conceptuelle des cours de médecine de l'Université Médicale Virtuelle Francophone (UMVF) », dans *Internet et Pédagogie Médicale 2003*, Faculté de Marseille.

Tran D.T et al. 2004a, « Lexicographie et recherche d'informations médicales par croisement de langues : une approche socioterminologique d'un lexique trilingue », dans *Journées d'étude Terminologie, Ontologie et Représentation des Connaissances*, ERSICOM - Université Jean-Moulin Lyon

Tran D.T et al, 2004b, « Experiments in cross-language medical information retrieval using amixing translation module », dans *MEDINFO 2004*, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pubmed

PRENDRE EN COMPTE LA DIMENSION GLOBALE D'UN CORPUS DANS LA CONTEXTUALISATION DU SENS : EXPERIMENTATIONS EN INFORMATIQUE LINGUISTIQUE

Pierre BEUST, Thibault ROY

GREYC CNRS UMR 6072 & pôle ModeSCoS
Université de Caen Basse-Normandie

Cet article s'inscrit dans le cadre de recherches en cours dans le domaine du Traitement Automatiques des Langues (TAL). Plus précisément, nous cherchons à mettre en œuvre des traitements sémantiques adaptés à certaines tâches informatisées où les spécificités socio-linguistiques des utilisateurs (par exemple leurs centres d'intérêt, leurs habitudes terminologiques) et leurs interprétations sont au centre de l'interaction homme-machine. C'est par exemple le cas dans bon nombre de tâches dans le domaine de la Gestion Electronique de Documents (GED) tels que le classement, l'archivage ou la recherche de documents. C'est plus encore le cas dans le domaine de la veille documentaire ou de la recherche d'information sur l'Internet.

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 travaillant avec ces documents et en produisant par ailleurs. Elle est centrale dans plusieurs phases de travail fréquentes que constituent l'élaboration et les analyses de corpus ou encore la construction et la gestion de ressources terminologiques. Nous allons montrer dans cet article comment nous envisageons et expérimentons dans nos travaux en informatique linguistique les rapports entre ces notions complexes et fortement connexes de sens, de contexte, de corpus et de ressources.

Après avoir posé notre cadre d'étude et plus précisément notre démarche centrée sur les besoins d'un utilisateur, nous décrirons comment nous abordons à travers nos recherches la question de la construction du sens. A partir de notre point de vue sur la nature du sens nous aborderons les notions de contexte, co-texte et d'intertexte. Nous préciserons par là ce que nous entendons par les rapports entre le local et le global. Nous présenterons alors nos travaux et nos expériences en cours dans le domaine de la cartographie thématique de corpus. Enfin nous ferons état des perspectives de recherche qui s'ouvrent à nous à ce moment de nos travaux.

Cadre d'étude

Les recherches que nous menons au sein de l'équipe ISLanD (Interactions, Sémiotique, Langues et Diagrammes) du laboratoire GREYC CNRS UMR 6072 à l'Université de Caen –

Basse Normandie s'inscrivent et trouvent leurs principales applications dans le cadre d'étude de la veille documentaire et de la recherche d'information, le plus généralement sur Internet. Afin de mettre en évidence les enjeux scientifiques, nous allons ici dresser un bref «état des lieux » des rapports entre ce cadre d'étude et de la problématique de la construction du sens.

Les technologies de l'information sur l'Internet forment un domaine d'application direct de l'ingénierie linguistique et plus précisément de l'accès au contenu des documents, d'où le rapport avec la problématique du sens. La taille des données textuelles à traiter ainsi que le nombre et la variété des traitements à réaliser rendent incontournable le développement de méthodes d'analyses automatiques fiables et rapides. A titre d'exemple, on peut se rappeler que le fameux moteur de recherche Google indexe aujourd'hui environ 8 milliards de documents et estimait en février 2003 traiter environ 250 millions de requêtes par jour.

Plusieurs types d'outils de TAL sont spécifiquement dédiés à la problématique du document. Ils constituent une évolution majeure du TAL aujourd'hui. Dans certains cas y sont réinvestis des travaux sur la compréhension des textes provenant de la tradition logico-grammaticale (c'est par exemple le cas des systèmes mis en compétition dans le cadre des conférences MUC[35]). Dans d'autres cas, on observe des démarches plus pragmatiques qui tentent de tirer de profit de larges corpus et de méthodes d'apprentissage automatiques (Claveau, 2003).

Adeline Nazarenko dans (Condamines & al., 2005, Chap. 6) établit quatre familles de méthodes automatiques d'accès au contenu des documents : l'extraction d'information, les méthodes de question/réponse, le résumé automatique et l'aide à la navigation. On entend par extraction d'information les méthodes qui consistent à rechercher dans un corpus très homogène (par exemple des dépêches d'actualité dans ou encore des articles scientifiques) des informations dont on sait qu'elles s'y trouvent. Ainsi on cherche par exemple dans un corpus d'actualité boursière à extraire les transactions de rachats et de fusions de sociétés ce qui revient à chercher à remplir des sortes de formulaires électroniques indiquant notamment qui a acheté qui, à quel prix et quand. Il est donc souvent visé ici d'alimenter de manière entièrement automatique des bases de données préexistantes à partir de corpus soigneusement sélectionnés. Les méthodes dites de Questions/Réponses n'ont pas le même objectif. Elles consistent à chercher un fragment de texte extrait d'un corpus volontairement assez généraliste dans lequel un sujet interprétant a de bonnes chances de trouver la réponse à une question qu'il aura formulée en langue naturelle. Par exemple extraire une séquence du style « *(…) la vie de Baudelaire, auteur des Fleurs du mal, fut (…)* » à la question « *Qui a écrit les Fleurs du mal ?* ». La bonne construction linguistique de la réponse n'est pas ici visée car il ne s'agit que de fournir une « fenêtre » dans une chaîne de caractères, éventuellement en essayant tout de même de ne pas couper des mots en leur milieu. Lors des conférences d'évaluation TREC9[36], les systèmes de questions/réponses avaient pour consigne de rendre des réponses de moins de 250 caractères à partir de 980 000 documents et de 700 questions. A la différence des méthodes d'extraction d'information, la construction de sens dans le cours de l'analyse est moins primordiale dans la mesure où finalement on s'en remet à l'interprétation d'un sujet humain. Les méthodes de résumé automatique s'appuient aussi largement sur l'interprétation de celui à qui est destiné le résumé. Bien souvent il est plus juste de parler de condensation ou de réduction de textes plutôt que de résumé (dans le sens de ce qu'est un résumé quand il est rédigé par un sujet humain). L'enjeu technique est de rechercher des phrases dont on pense qu'elles ont un statut assez significatif (par exemple une phrase qui commencerait par « *en somme, on constate que (…)* » a de bonnes chances de synthétiser ce qui est dit avant) et de les juxtaposer dans un «résumé » où l'on compte que

[35] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/proceedings_index.html (consultée le 29-06-2006)

[36] http://trec.nist.gov/pubs/trec9/t9_proceedings.html (consultée le 29-06-2006)

celui qui le lira pourra rétablir une certaine cohérence textuelle, par exemple relativement aux rattachements anaphoriques.

Les méthodes d'extraction d'information, de Question/Réponse et de résumé automatique s'adressent principalement à la dimension rhématique des documents en cherchant d'une certaine façon à savoir ce qui est dit, où et comment. En général, les méthodes d'aide à la navigation s'adressent plus spécifiquement à la dimension thématique des documents (dans le sens où l'on cherche de manière plus globale à savoir de quoi traite un document ou un ensemble de documents). Les applications les plus courantes de ces méthodes sont l'indexation de document, l'extraction de terminologies, l'aide à la lecture (visualisation de documents ou encore création d'index par exemple), le groupement en classes de documents, la cartographie de corpus (qui, comme nous le verrons par la suite, est l'objet de nos recherches et nos développements).

Les quatre familles de méthodes d'accès au contenu présentées ci-dessus regroupent des projets de recherche où sont mis en œuvre beaucoup d'intelligence du point de vue des collaborations interdisciplinaires, notamment entre la linguistique et l'informatique. Cependant force est de constater que la majeure partie de ces projets de recherche sont toujours à l'état de prototypes de laboratoire et sont jusqu'à présent très peu mis en application et évalués dans des outils sur Internet à destination du plus grand nombre. Cela a des conséquences comme le montrent (Lavenus & al., 2002) à propos des méthodes de Question/Réponse en mettant en évidence la différence entre les corpus de référence utilisés dans les conférences TREC par rapport à des vraies questions d'utilisateur en recherche documentaire. Les auteurs notent que les questions du corpus de référence sont toutes des interrogatives canoniques courtes (par exemple *« What does a defibrillator do ? »*) alors que la majorité des demandes de « vrais » utilisateurs sont couramment des affirmatives complexes du style « *je voudrais savoir (…)* ».

Paradoxalement, si d'un point de vue informatique et algorithmique les méthodes couramment utilisées notamment par les moteurs de recherche sont très fines et fiables, on constate effectivement qu'elles restent linguistiquement très pauvres, à la fois du point de vue de leur fonctionnement propre mais également du point de vue de l'interaction avec leurs utilisateurs.

Les méthodes d'indexation utilisées par les moteurs de recherche pour associer des documents à des mots clés potentiels en sont un exemple. Cette indexation est dite « *Full Text* » dans le sens où tous les mots figurant dans un document sont gardés comme entrée d'index pour ce document. Pas étonnant dans ces conditions que les mots grammaticaux indexent une multitude de documents (expérience faite sur Google le 25/8/05 : une recherche stupide avec le mot clé unique « de » donne 873 000 000 réponses[37] et encore il est clair qu'en ce qui concerne les mots grammaticaux et le nombre de réponses possibles potentiellement, seules les pages considérées comme relativement importantes sont rendues). Ceci a des inconvénients, notamment la taille énorme des bases d'index que le moteur de recherche doit archiver et doit être capable d'interroger rapidement. En fait, l'intérêt de cette indexation un peu brutale réside dans le fait de pouvoir garder facilement comme index tout ce qui dans un texte ne peut être retrouvé dans un dictionnaire. C'est surtout le cas des entités nommées telles que des noms propres, des expressions temporelles ou encore des noms de quantité qui restent importants par rapport aux documents (on imagine mal par exemple que le nom d'une

[37] Il ne faut rester assez prudent sur l'interprétation du nombre de réponses rendu par un moteur de recherche car ce nombre est en fait souvent estimé plutôt que réel. Cette estimation en parfois fausse de manière évidente comme en témoigne une expérience menée par Jean Véronis (cf. http://aixtal.blogspot.com/2005/01/web-google-perd-la-boole.html (consultée le 29-06-2006)) en février 2005 où la requête Chirac rendait 3,2 millions de documents alors que la requête Chirac OR Sarkozy en rendait un peu moins de 2 millions, ce qui n'est pas logiquement cohérent.

société ne puisse pas être gardé comme entrée d'index pour son site web) mais dont il est difficile de dresser un catalogue fiable et durable[38]. La question du repérage et même de l'étiquetage (en tant que nom d'organisation ou de nom de lieu par exemple) des entités nommées est un enjeu important du TAL aujourd'hui et de nombreux projets de recherche abordent cette question avec des résultats intéressants mais leurs avancées n'ont pas encore eu de retombées sur les méthodes d'indexation utilisées sur Internet.

Dans leurs interactions avec les utilisateurs, les moteurs de recherche sont souvent assez rudimentaires d'un point de vue linguistique. Il faut bien souligner que l'utilisateur et son objectif de recherche sont uniquement considérés sous la forme d'une liste de mots clés (dont la casse et l'accentuation et même l'ordre sont d'ailleurs rarement pris en compte[39]) considérés pour une seule recherche dans la mesure où toutes les requêtes sont traitées indépendamment les unes des autres. Dans la pratique on s'aperçoit que pour mener à bien une recherche sur le web, il convient en fait d'interroger successivement plusieurs fois le (ou les) moteur(s) en ajoutant ou en précisant certains mots clés en fonction des résultats rendus à chaque étape. C'est donc le plus souvent à l'utilisateur seul qu'il convient de développer des stratégies efficaces pour trouver des mots clés adaptés à sa recherche. Certaines tentatives sont mises en place par certains moteurs pour aller un peu plus loin que la simple prise en compte de mots clés. Par exemple, Google permet de rechercher un mots clé ou un de ses synonymes avec l'opérateur tilde ~ (par exemple une recherche sur `powerpoint ~help` effectuera une recherche sur `powerpoint ET help ou tips, faq, tutorial`). Cependant, c'est le moteur lui-même qui établit ses listes de synonymes et il serait peut être plus judicieux que celles-ci soit validées par les utilisateurs quand ils les utilisent. Il convient donc, en tant qu'utilisateur, de rester très prudent quant aux compétences linguistiques des moteurs. Toujours à propos de Google, on trouve un exemple de résultat assez malheureux de l'opérateur `define` sur le *blog* de Jean Véronis[40]. L'opérateur `define` (disponible pour les pages en français depuis avril 2005) sert à rechercher à propos d'un mot des pages Web où ce mot ferait visiblement l'objet d'une définition. L'expérience relatée consiste à rechercher ainsi sur Google une définition du mot *femme* avec la requête `define:femme`. Les résultats donnés sont pour le moins plus que contestables. On aurait donc bien tort de croire à la fiabilité de l'opérateur `define` (qui pourtant est présenté par Google comme un outil de recherche de définition sans plus de détails) comme on aurait tort aussi de considérer le Web dans son ensemble comme une encyclopédie dans lequel on puisse rechercher des définitions attestées, notamment d'un point de vue moral.

En matière d'ingénierie documentaire la tendance actuelle est pourtant de renforcer ce genre d'utilisation du Web en cherchant à en faire une vaste base de connaissances, ce qu'évidemment il n'est pas. C'est la démarche considérée dans le projet du Web Sémantique où l'objectif annoncé par Tim Berners-Lee (Berners-Lee, 1998), initiateur du projet et directeur du W3C, est d'enrichir (notamment au moyen des technologies développées autour du langage XML) les documents (à l'aide d'ontologies normalisées, soit automatiquement, soit en assistant leur auteurs) avec des informations sur leur propre sémantique qui soit directement interprétables par des agents logiciels sans la supervision d'une interprétation

[38] Bien qu'elles soient délicates à construire et maintenir dans le temps, de telles ressources existent. C'est le cas de CELEX (CELEX, 1998), un lexique de 160 595 mots fléchis (avec leur lemme et leur catégorie syntaxique), une liste de 8 070 prénoms et de 211 587 noms de familles, une liste de 22 095 entreprises et 649 noms d'organisations, une liste de 7 813 villes et une autre de 1 144 pays et une liste sur les unités physiques et monétaires. Certaines de ces listes proviennent de sources déjà connues (c'est le cas des 22 095 entreprises issues du « *Wall Street Research Network* »), d'autres (les noms d'organisations) sont issues d'une acquisition lexicale sur Internet (Jacquemin & al., 2000) et d'autres sont construites manuellement (par exemple les unités physique et les monnaies)

[39] C'est par exemple le cas du moteur Exalead : http://www.exalead.com (consultée le 29-06-2006)

[40] http://aixtal.blogspot.com/2005/04/web-la-femme-selon-google.html (consultée le 29-06-2006)

humaine. Ceci fait l'hypothèse que la valeur sémantique d'un passage de document est le fait de son auteur alors que c'est finalement bien plus celui de son lecteur. L'expérience sur la définition de *femme* nous apprend bien qu'une définition considérée comme telle par quelqu'un n'a pas pour autant cette valeur pour d'autres et qu'au final c'est celle de l'utilisateur du moteur qu'il faudrait considérer.

Notre approche de la veille documentaire sur l'Internet se situe à l'opposé de celles défendues dans le cadre du Web Sémantique. Elle s'en distingue essentiellement par le fait que nous mettons l'accent sur des traitements et des ressources termino-ontologiques (bases de données terminologiques, représentations du contenu lexical etc.) avant tout centrés sur leur utilisateur, de sa tâche, de ses besoins et de ses centres d'intérêt. Là où le Web Sémantique cherche à rendre le plus possible partagées de vastes ontologies qui synthétisent une connaissance pensée comme objective et devant convenir à tous les utilisateurs, nous préférons manipuler des ressources propres à un utilisateur ou un petit groupe d'utilisateurs. Il en découle une certaine *légèreté* de ces ressources, au sens de (Perlerin, 2004), dans la mesure où elles ne représentent que ce qui est important du point de vue de l'utilisateur et restent ainsi de taille raisonnable (par exemple une centaine de termes) ce qui les rend moins complexes à construire, à maintenir et à enrichir.

Cette approche centrée utilisateur conduit à opérer un certain renversement scientifique relativement aux ressources qu'utilisent les modèles de TAL. Premièrement, d'un point de vue très pratique, force est de constater que des ressources très généralistes, valables pour tout type de traitement envisagé ainsi qu'à destination de tout utilisateur potentiel, ne sont pas facilement disponibles (sous forme électronique pour des traitements automatiques) et encore moins gratuites. Deuxièmement, nous soutenons que l'idée même d'une ressource généraliste est illusoire car elle dépend inévitablement du contexte qui lui préexiste (le but recherché par le ou les auteurs ainsi que leurs spécificités socioculturelles). Le rapport de l'Action Spécifique 32 du CNRS/STIC en 2003 (Charlet & al., 2003) va également dans ce sens en précisant un obstacle au projet du Web Sémantique : la détermination et l'ajout, même de simples méta-données, n'est pas une activité naturelle pour la plupart des personnes.

La tradition logico-grammaticale et plus précisément la sémantique formelle et computationnelle cherchent à représenter et à produire, automatiquement ou pas, des formes le plus possible objectivées des significations et du sens. Dans la démarche centrée utilisateur, on part d'une position duale où l'on considère que les traitements sémantiques appliqués à l'accès au contenus des documents ont tout à y gagner à être le plus possible subjectivés, tant du point de vue des ressources que du point de vue des résultats opératoires. Cette démarche nous paraît être une réponse au constat que dressent Didier Bourigault et Nathalie Aussenac-Gilles à propos de la variabilité des terminologies :

> *(…) le constat de la variabilité des terminologies s'impose : étant donné un domaine d'activité, il n'y a pas* une *terminologie, qui représenterait le savoir du domaine, mais autant de ressources termino-ontologiques que d'applications dans lesquelles ces ressources sont utilisées (Bourigault & al., 2003).*

Les ressources utilisées dans nos expérimentations sont produites de manière endogène dans une boucle d'interaction entre un outil logiciel, un utilisateur et des corpus où chaque pôle est déterminant. Il en découle une importance significative des corpus utilisés qui du coup ne peuvent plus être considérés uniquement comme un réservoir de formes attestées sur lequel on tenterait de mettre en œuvre un calcul à base de ressources exogènes. Le corpus utilisé dans le cadre de nos outils de TAL est à l'origine des ressources lexicales construites et constitue en même temps le matériau d'expérimentation de nos propositions. Ainsi notre démarche s'inscrit dans un processus de recherche et de développement en aller-retour entre des outils (des logiciels d'étude), des corpus (des corpus d'étude) et des utilisateurs, les uns étant conditionnés par les autres.

Comme on le voit bien, bon nombre de démarches diffèrent grandement dans les méthodes d'accès au contenu. Il y a un point incontournable sur lequel il convient également de préciser les ancrages épistémologiques, c'est la notion même de contenu et plus largement la question du sens. C'est ce que nous allons faire très succinctement dans la partie suivante.

La question de la construction du sens

Dans une communication sur l'histoire des traitements sémantiques en TAL, Gérard Sabah[41] précise ce que peut être le sens du point du résultat visé par telle ou telle sémantique :

- préciser les conditions de vérité de l'expression traitée (dans le cas d'une sémantique véri conditionnelle) ;
- décrire une expression comme l'ensemble des propriétés théoriques que possèdent les concepts correspondants (comme en sémantique intensionnelle) ;
- décrire une expression comme l'ensemble des objets ou des situations du monde de référence que cette expression peut désigner (on parlera alors de sémantique extensionnelle ou également de sémantique dénotationnelle ou référentielle) ;
- chercher à décomposer le contenu des mots en éléments de sens plus primitifs pour étudier les possibilités de combinaison de ces éléments (on est ici dans le cadre d'une sémantique componentielle) ;
- décrire une expression comme l'ensemble des actions à effectuer pour trouver l'objet désigné (on parle ici de sémantique procédurale) ;
- mettre en évidence les marqueurs et les constructions utilisées pour qu'un énoncé puisse servir comme un argument en faveur d'un autre énoncé (il s'agit alors d'une sémantique argumentative).

Cette catégorisation, qui ne se veut pas exhaustive, montre la grande variété des points de vue sur le sens. Aucune de ces sémantiques n'a complètement tort ou complètement raison. Par exemple, il existe bon nombre d'énoncés en langue naturelle dont le sens a un certain rapport avec la vérité ou encore la référence, mais cela ne veut pas dire que *tout* énoncé est interprétable en terme de vérité de ou de référence. Il faut donc en adoptant l'une ou l'autre de ces sémantiques garder à l'esprit que l'on ne cherchera avec à rendre compte uniquement que d'une partie du sens qui de fait ne le couvre pas dans son ensemble.

Selon la catégorisation de Gérard Sabah, nous situons notre approche de la notion de sens dans le cadre d'une sémantique componentielle et plus précisément dans un héritage scientifique de la Sémantique Interprétative (SI) de François Rastier (Rastier, 1987), elle-même en filiation avec les travaux en sémantique structurale dont l'origine remonte à Hjelmslev. De la SI on retient principalement deux aspects. Premièrement on en tire (parfois avec quelques adaptations au cadre d'une instrumentation informatique[42]) un « appareillage » théorique fin pour la description d'effets de sens. Par exemple, les notions de sèmes, d'isotopies (*i.e.* récurrences d'un même sème dans un texte), etc. Deuxièmement, on s'approprie de la SI un positionnement épistémologique relativement à la question du sens. Celui-ci consiste à préférer la tradition rhétorique et herméneutique à la tradition logico-grammaticale. Ainsi on défend, d'une part, le principe selon lequel le global détermine le local (ce qui marque une rupture avec le principe de compositionnalité) et, d'autre part, que le sens ne peut pas être intégralement objectivé (Rastier, 1998).

Si le sens ne peut pas être objectivé, il peut encore moins l'être de manière formelle (comme cela est souvent le cas dans la tradition logico-grammaticale). On rejoint ici l'avis de (Nicolle, 2005) pour qui le sens n'est jamais capturé par ses représentations. Toute

[41] http://www.limsi.fr/Individu/gs/textes/ATALA-14.12.96/LePointSurLeSens.html (consultée le 29-06-2006)
[42] C'est par exemple le cas concernant la notion de sème que nous avons redéfini dans (Beust, 1998).

représentation du sens est forcément incomplète et il n'y a donc pas de langage formel qui puisse reproduire fidèlement le sens d'un énoncé en langue naturelle alors que tout énoncé formel peut être reformulé dans une langue. Anne Nicolle en tire la conséquence que la langue est un langage terminal. L'interprétation langagière est en cela bien distincte de l'interprétation logique qui se résume à la traduction dans un autre langage. Dans le cas de l'interprétation langagière, il n'y a pas d'autre langage.

Ainsi, à la manière de Jacques Courcil qui définit les principes de non consignation et de non préméditation de la chaîne parlée (Courcil, 2000), nous défendons un principe de non transformation du sens en langue naturelle dans la mesure où il n'y a pas de pensée construite possible qui ne soit pas déjà sous forme langagière. Le sens est donc une réalité concrète intralinguistique et subjective. Dans le cadre de la SI, l'interprétation est considérée comme une perception sémantique, perception forcément individuelle, dont toute tentative d'objectivation est une sommation incomplète de points de vue. Ainsi le sens d'un texte est une interprétation à un moment donné et dans une tâche donnée d'un sujet interprétant, ce qui est à notre avis un argument fort pour une instrumentation de la sémantique des langues individu-centrée.

Beaucoup de travaux en sémantique formelle (logique, DRT[43], SDRT[44] etc.) ont depuis des années déployé beaucoup d'intelligence pour obtenir de façon compositionnelle un « calcul du sens » acceptable. Force est de constater qu'un tel résultat n'est toujours pas atteint à l'heure actuelle. Il ne s'agit pas ici que d'un problème d'évaluation dont on n'aurait pas encore bien mis en place la méthodologie mais d'un problème beaucoup plus profond. Dès lors qu'on parle de « vrais » textes, de « vrais » corpus, et pas simplement de phrases d'exemples artificiellement construites en dehors d'un contexte linguistique et pragmatique, il convient de se rendre compte que la dimension interprétative personnelle fait qu'il n'y a pas de consensus évident sur ce qu'est ou n'est pas le sens d'un texte. Il en résulte, à notre avis, que le sens ne peut être modélisé à la façon d'un résultat calculatoire qui serait plus ou moins complété ou dégradé d'un interprétant à un autre. Le sens n'est pas de nature symbolique ; c'est un processus sémiotique au centre de l'activité de l'interprétant qui est complexe, notamment parce qu'il est réflexif.

Dès lors, la construction du sens n'est pas une question d'extraction à partir du matériau linguistique, ni même de calculs sur une extraction d'informations à partir du matériau linguistique. Ce ne sont pas tant les caractéristiques propres des mots, des phrases ou des paragraphes qui priment dans le sens des textes mais c'est ce que les interprétants en attendent ou y projettent. Des critères externes aux textes peuvent être tout aussi importants. D'une certaine manière, le succès du moteur de recherche Google nous en donne un exemple à propos du rapport entre le contenu d'une page et l'importance de cette page du point de vue des utilisateurs du moteur de recherche. Comparativement à d'autres moteurs de recherche, le classement par importance des pages Web répondant à une requête ne dépend pas avant tout de leur contenu. Pour AltaVista par exemple, la pertinence d'une page dépend de critères liés à son contenu (présence répétée d'un mot clé de la requête dans le contenu, dans le titre ou dans les méta-données). Pour Google, c'est l'algorithme de *PageRank*[45] qui conditionne la pertinence d'une page avec, en plus des techniques classiques donnant une importance particulière à certaines zones (par exemple les titres), le principe suivant[46] : plus il existe de pages qui ont un lien vers la page P, plus P est pertinente (quelle que soit son contenu et quels que soient les mots clés de la requête). Les travaux menés dans le cadre du projet PRINCIP (Valette & Grabar, 2004) visant la détection automatique de sites Internet au contenu illicite

[43] Discourse Representation Theory (Kamp, 1981)
[44] Segmented Discourse Representation Theory (Asher, 1993)
[45] http://www.google.com/technology/ (consultée le 29-06-2006)
[46] Par rapport à d'autres moteurs de recherche, ce principe donne à Google un caractère résolument socio-centré.

(principalement des propos racistes ou antisémites) montrent également que le sens est de nature pluri-sémiotique. Il provient de la présence conjointe de plusieurs facteurs dont certains sont extérieurs aux textes. Ainsi, du point de vue de la thématique des textes, des propos racistes ou anti-racistes sont parfois très proches à tel point qu'une détection automatique fiable uniquement basée sur la recherche de certains mots clés du texte n'est pas facile à obtenir. Si en plus on veut qu'elle soit fiable dans la durée, cela devient très difficile étant donné qu'on ne peut pas prévoir à l'avance l'usage de certains mots dans certains contextes. Il faut alors exploiter d'autres critères pour fiabiliser cette détection : la ponctuation qui traditionnellement n'est pas prise en compte (on remarque de façon statistiquement significative que les sites racistes utilisent fortement le point d'exclamation et que les sites anti-racistes utilise plutôt des points de suspension), le type de police de caractère utilisé (la police Arial semble significativement caractéristique des sites racistes), les couleurs de fond et de police de caractères utilisées (le rouge et le noir sont aussi significativement caractéristiques des sites racistes), les contenus des images entourant le texte (la thématique de l'animal dans les textes racistes est souvent corrélée avec des dessins montrant des animaux souvent connotés de façon péjorative, le rat par exemple).

Dans la logique de nos partis pris épistémologiques, la question de la construction du sens se trouve déplacée. Nous ne l'abordons pas avec l'idée d'un traitement automatique qui produirait un résultat (quand bien même ce résultat serait un processus). Nous l'abordons d'une autre manière sous l'angle de l'instrumentation informatique dédiée à *l'assistance* à l'interprétation. Sous cet angle de vue il convient de développer des outils logiciels pour mettre en place des interactions homme-machine où l'utilisateur, les textes et ses interprétations sont l'objet de la boucle interactive. L'adaptation des interfaces à cette boucle interactive, notamment par l'utilisation de méthodes de visualisation adéquates tient une place importante (c'est notamment ce qu'ont bien compris les terminologues qui ont travaillé à la réalisation de concordanciers). On va donc chercher à développer des méthodes d'accès au contenu qui plutôt que d'extraire du sens vont chercher à alimenter des interactions (notamment des techniques de visualisation telles que des cartes ou des diagrammes) avec des signes qui ont pour objectif de faire sens du point de vue de l'utilisateur. La mise en œuvre de ces méthodes d'accès au contenu confère, comme nous allons le voir dans la suite, un statut important à la dimension globale des corpus.

Le rapport local/global dans la contextualisation du sens

Le principe de détermination du local par le global propre à l'approche herméneutique est un principe de contextualisation du sens (au passage, principe alternatif à la compositionnalité). La question de la contextualisation du sens de telle ou telle unité linguistique du texte (mot, syntagme, paragraphe par exemple) est d'une grande importance dans la perspective de la sémantique interprétative car elle est la base de l'établissement de parcours interprétatifs, c'est-à-dire des suites d'opérations permettant d'assigner un ou plusieurs contenus à des expressions (ce qui explique comment les langues peuvent s'acquérir réflexivement par leur pratique).

Dans la contextualisation entrent en compte, selon nous, 3 notions : le co-texte, le contexte extralinguistique et l'intertexte :

- On entend par co-texte d'une unité linguistique son « entourage » dans le texte, c'est-à-dire un passage de texte : une zone de localité sémantique pertinente autour d'une unité. Cette zone est appelée Période (Rastier & al. 1994, p. 116) et elle est délimitée par l'étendue des relations d'isotopies, de prédication et d'anaphore ;

- Le contexte extralinguistique regroupe les conditions pragmatiques liées à l'interprétation du texte. Dans le cadre de nos expérimentations en recherche d'information, on limitera ce contexte à l'utilisateur et sa tâche (voire au groupe d'utilisateurs et leur tâche) ;
- L'intertexte rassemble tous les documents que l'utilisateur estime liés à un texte du point de vue de son interprétation. Tout texte mis en relation avec d'autres textes en reçoit des déterminations sémantiques et modifie potentiellement le sens de chacun des autres textes (c'est le principe d'architextualité définit dans (Rastier, 2001). Un document peut appartenir à plusieurs intertextes comme un texte peut s'interpréter dans plusieurs intertextes en fonction des relations sémantiques établies, c'est-à-dire des objectifs interprétatifs. On peut penser par exemple qu'un même article de presse n'indique pas le même point de vue quand il y est fait référence dans une revue de presse des plus sérieuses ou dans les colonnes d'un journal satirique.

Ainsi contextualiser, c'est établir au sein du co-texte des parcours interprétatifs qui tiennent compte du contexte extralinguistique et de l'intertexte. Ce que l'on peut analyser avec les moyens d'une sémantique componentielle comme mettre en évidence dans certains passages du texte des sèmes particulièrement importants. Ainsi l'analyse de la détermination du local par le global consiste à identifier localement des sèmes pertinents issus du global (le contexte ou l'intertexte). Le contexte et l'intertexte restent deux notions aux contours assez flous qu'il convient de circonscrire ici. Comme nous le verrons par la suite, le contexte est vu dans le cadre de nos expérimentations via les ressources personnelles construites par le ou les utilisateurs et l'intertexte via le corpus d'étude[47] sur lequel les outils logiciels développés permettent de travailler. A l'égal des ressources exploitées, le corpus « matérialisant » l'intertexte en tire un rôle central par rapport à la construction du sens en contexte, ce qui lui confère bien plus d'importance qu'un simple réservoir de formes attestées.

Identifier et caractériser l'importance d'un sème dans la contextualisation du contenu d'une unité est le résultat d'opérations interprétatives d'actualisation et de virtualisation de sèmes. Dans l'énoncé *Le facteur m'a donné une lettre* le sème /courrier/ est actualisé dans le contenu de *lettre* parce qu'il se répète dans le contenu de *facteur* formant ainsi une isotopie. Cette actualisation permet de retenir la signification pertinente de *lettre* dans l'énoncé (on ne retient donc pas, par exemple, la signification de *lettre* en tant que caractère de l'alphabet) et précise une sélection du co-texte sur une partie du signifié de *lettre*. Ainsi, dans cet exemple, le sème /courrier/ est renforcé par le co-texte alors que ce n'est notamment pas le cas du sème /en papier/ appartenant également au contenu de *lettre* ; à l'inverse, ce sème serait probablement actualisé dans *Il a chiffonné sa lettre* et pas /courrier/. La virtualisation est l'opération interprétative duale de l'actualisation. Elle décrit une neutralisation d'un sème en contexte. Par exemple, dans le syntagme *Le chat immortel*[48], on dira que le sème /mortel/ appartenant au contenu de *chat* est virtualisé car non seulement il n'est pas répété dans l'énoncé mais, de plus, il est invalidé par le contenu sémique de *immortel*. La virtualisation ne doit pas être considérée comme un simple retrait d'un sème. L'idée d'une neutralisation temporaire est plus juste car, si dans la suite du texte et/ou de l'intertexte le sème virtualisé venait à réapparaître dans d'autres unités, il serait alors ré-actualisé.

L'actualisation et la virtualisation jouent un rôle important sur la mise en co-texte de contenus sémiques définis en langue, c'est-à-dire sur les sèmes dits inhérents. Cette notion de sème inhérent est à opposer à celle de sème afférent. Dans des contextes particuliers, on peut

[47] Le corpus n'est pas pour autant une simplification de l'intertexte, c'en est selon (Rastier 1998, note de bas de page n°17) une objectivation : « *Le corpus est la seule objectivation possible (philologique) de l'intertexte, qui sinon demeure une notion des plus vagues* ».

[48] Extrait de la phrase «*Le chat immortel a fui sur les toits comme s'il avait un démon à ses trousses.* » sur la page web http://rouscaille.tripod.com/rouscaill/id163.html (consultée le 29-06-2006).

actualiser des sèmes qui ne font pas partie des contenus des unités du texte. Ces sèmes sont dits afférents et l'opération interprétative qui consiste à les actualiser porte le nom d'afférence. L'afférence consiste en la production d'un sème qui vient créer ou renforcer une isotopie. Selon Thierry Mézaille[49], il y a une telle production d'un sème isotopant dès lors qu'est établie une relation sémantique d'assimilation ou de dissimilation. L'assimilation est une afférence co-textuelle (c'est-à-dire que le sème afférent est inhérent dans d'autres unité du co-texte) d'un sème générique par présomption d'une récurrence sémique. Par exemple, dans l'énoncé *Voici des choux, des concombres et des scoubidous* l'assimilation consiste en une afférence d'un sème à *scoubidou* où le but est de renforcer une répétition déjà initiée dans le co-texte. En l'occurrence, il s'agit d'enrichir le contenu sémique de *scoubidou* avec le sème afférent générique[50] /légume/ pour rendre le thème de l'énoncé uniforme[51]. L'opération interprétative inverse de l'assimilation est la dissimilation. Alors que l'assimilation diminue, par afférence, les contrastes forts, la dissimilation, quant à elle, augmente les contrastes faibles. La dynamique des sèmes en cause dans la dissimilation n'est plus une afférence de sème générique, comme c'est le cas pour l'assimilation, mais une afférence de sèmes spécifiques pour différencier, dans un co-texte, les contenus sémantiquement proches. Par exemple, dans l'énumération *routes et autoroutes*, le contenu de *route* doit décrire une signification spécifique qui exclut la signification de *autoroute* et non une signification générique qui inclut cette signification. La dissimilation est encore plus flagrante dans l'exemple suivant où il y a répétition de la même lexie dans *Il y a musique et musique*. Ici, la dissimilation consiste, à distinguer par des sèmes spécifiques les deux significations de *musique*. Ainsi, on peut afférer à la première le sème spécifique /agréable/ et à la seconde /désagréable/.

L'assimilation et la dissimilation sont des formes d'afférences co-textuelles mais la sémantique interprétative décrit également une autre forme d'afférence : l'afférence socialement normée. Dans de tels cas d'afférence, il y a bien enrichissement contextuel du contenu d'une lexie dans un énoncé par un (ou plusieurs) sème(s) (c'est bien pour cela qu'il s'agit toujours d'une afférence) mais cette fois, le ou les sèmes afférents ne sont pas inhérents dans d'autres contenus d'unités linguistique du co-texte. L'afférence est alors le fait d'une norme sociale partagée au sein d'une communauté linguistique. C'est, par exemple, le cas du sème /tristesse/ afférent au contenu de *noir* dans *il broie du noir* ou encore le cas du sème /bonheur/ dans *rose* dans *la vie en rose*. Là où l'afférence co-textuelle (par assimilation ou dissimilation) est le résultat d'un parcours interprétatif local, l'afférence socialement normée résulte de parcours interprétatifs beaucoup plus globaux (à titre d'exemple d'afférences globales sur l'ensemble d'un corpus, on peut citer l'étude de (Rastier, 1987) sur le roman de Stendhal *Le rouge et le noir* où il y a une afférence jusque dans le titre du roman des sèmes /armée/ et /Église/).

Le concept d'afférence est un outil théorique très fin pour la description de la dynamique des sèmes dans les corpus et pour la caractérisation des effets du global sur le local. Le problème de la modélisation de l'afférence (surtout en ce qui concerne l'afférence socialement normée) c'est qu'il faudrait avoir des ressources très larges pour pouvoir l'expérimenter de façon opératoire dans un traitement automatique. Comme on l'a dit plus

---

[49] cf. « Quels mécanismes pour (r)établir la cohésion sémantique textuelle ? Sur la prééminence des processus d'assimilation et de dissimilation dans l'interprétation des énoncés contradictoires et métaphoriques » disponible en ligne à l'adresse : http://www.chez.com/mezaille/contraphore.htm (consultée le 29-06-2006).

[50] L'assimilation concerne bien des sèmes génériques car c'est le sème /légume/ qui est afférent et pas le sème /vert/ inhérent à *choux* et *concombre* mais spécifique, ce qui a pour conséquence qu'il n'y ait pas de défaut d'assimilation dans *Voici des choux, des concombres et des carottes*).

[51] On retrouve ce type d'assimilation dans certaines formes d'humour lorsque le sème afférent est particulièrement inattendu dans le contenu sémique de la lexie en cause. Le titre du livre de G. Lakoff : *Women, Fire and Dangerous Things : What Categories Reveal about the Mind* (1987) en témoigne en provocant l'afférence du sème /dangereux/ au lexème *Women*.

haut, un texte n'est pas lié à un unique intertexte. Du point de vue du sujet interprétant et de son histoire propre, tous les intertextes forment un univers de textes construit individuellement de manière le plus souvent inconsciente. Théodore Thlivitis l'appelle l'*anagnose* (Thlivitis 1998, p. 41). Dans la perspective centrée utilisateur qui est la nôtre (ainsi que celle Théodore Thlivitis) l'anagnose serait idéalement ce qu'il faudrait formaliser pour représenter au mieux l'utilisateur et modéliser avec satisfaction ses afférences. Seulement, il n'est pas du tout évident, au contraire, que cette anagnose rassemblant l'histoire d'un individu (l'histoire de ses interprétations comme son histoire propre), sa culture, sa société, les données de son époque soit formalisable. On est ici face aux mêmes problèmes que ceux de la constitution des ontologies généralistes. Dés lors, les différentes ressources personnelles que nous serons amenés à construire et à manipuler ne toucheront que des infimes parties de cette anagnose. Nous ne prétendons donc pas rendre compte de l'ensemble des mécanismes d'afférence dans nos expérimentations informatiques. Ce n'est pas pour autant un problème majeur car il est possible de personnaliser efficacement (comme l'a montré Vincent Perlerin) des tâches de recherche d'information et d'accès au contenus de documents avec des ressources légères car d'emblée non exhaustives. C'est ce que nous allons montrer dans la suite en détaillant nos expérimentations à propos de la cartographie thématique.

La cartographie thématique : expérimentation sur corpus

La plate-forme ProxiDocs[52] (Roy et Beust, 2004) permet de construire différentes représentations globales d'un corpus de textes à partir de ressources terminologiques construites par l'utilisateur. Ces représentations sont appelées des cartes. Ce sont des visualisations topologiques interactives, personnalisées en fonction d'un utilisateur ou d'un petit groupe d'utilisateurs.

Les ressources utilisées pour produire les cartes vont représenter les thèmes ou les domaines choisis par l'utilisateur pour intervenir dans ses analyses. Deux types de représentations sont possibles selon les besoins de l'utilisateur : la représentation en «sacs de mots »où chaque thème est représenté par un ensemble de lexies s'y rapportant selon le point de vue de l'utilisateur ; et la représentation selon le modèle LUCIA de sémantique différentielle (Perlerin, 2004) où chaque domaine est représenté par un ensemble (appelé un dispositif) de catégories de lexies dont la signification est représentée par des différences de sèmes. Afin de construire de telles ressources terminologiques, nous proposons un ensemble de logiciels d'étude[53] complémentaires apportant une aide à l'utilisateur. Par exemple, les outils MemLabor (Perlerin, 2002) et FlexiConcord permettent une première analyse du corpus d'étude en réalisant respectivement une extraction des graphies répétées et une mise en contexte de termes et de leurs flexions. Après avoir isolé des termes pouvant intervenir dans ses analyses, l'utilisateur peut organiser ces termes en classes thématiques (qu'on appelle thèmes) à l'aide de l'outil ThemeEditor (Beust, 2002). Un principe de surlignage avec différentes couleurs (une couleur correspondant à un thème) permet de mettre en évidence la répartition, l'alternance et les enchaînements au long d'un texte des thèmes ainsi construits. Selon les besoins de sa tâche, l'utilisateur peut choisir de construire des représentations sémantiques plus fines selon le modèle LUCIA. Pour cela, l'outil LuciaBuilder (Perlerin, 2004, pp. 151-160) est mis à sa disposition afin de l'assister dans les différentes étapes de création des dispositifs représentant les domaines de son choix.

[52] http://www.info.unicaen.fr/~troy/proxidocs (consultée le 29-06-2006).

[53] Ces logiciels d'étude sont tous disponibles sur le site de l'équipe ISLanD du laboratoire GREYC de l'université de Caen : http://www.greyc.unicaen.fr/island/logiciel/ (consultée le 29-06-2006).

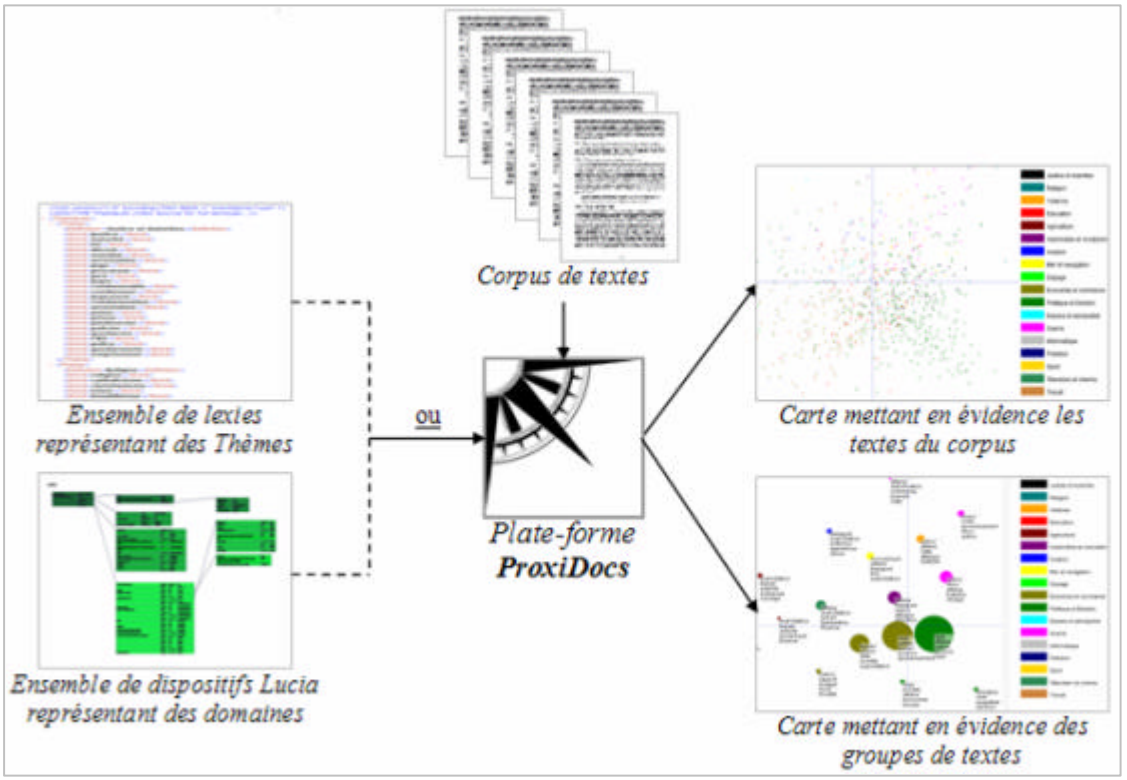

Figure 1 : Utilisation de la plate-forme ProxiDocs.

La plate-forme ProxiDocs permet d'opérer différentes visualisations d'une analyse globale du corpus à partir des ressources de l'utilisateur (la figure 1 illustre les possibilités de la plate-forme exploitées dans cet article) :

- des cartes en 2 ou 3 dimensions représentant chaque texte du corpus analysé par un point. Chacun de ces points est un lien hypertexte vers le texte où les lexies des thèmes ou des dispositifs sont mises en valeur à l'aide d'une technique de coloriage. La couleur d'un point correspond au thème majoritaire repéré dans le document représenté. De telles cartes sont utiles afin d'étudier les liens et les différences de thématiques abordées entre les documents du corpus ;
- des cartes en 2 ou 3 dimensions mettant en évidence des groupes de textes abordant des thèmes proches. Chaque groupe est représenté sur la carte par un disque ou une sphère de diamètre proportionnelle au nombre de textes qu'il contient. La couleur attribuée au disque ou à la sphère correspond au thème majoritaire repéré dans les textes du groupe. Les disques représentant les groupes portent un jeu d'étiquettes (au plus 5 étiquettes) indiquant les lexies localement les plus fréquentes dans l'ensemble des documents du groupe. Chacun de ces groupes est un lien hypertexte vers un rapport sur le contenu du groupe représenté. Ce rapport indique le texte le plus «représentatif » du groupe (celui situé le plus près du centre de gravité du groupe), présente la répartition des thèmes ou des catégories abordés, met en évidence les lexies répétées et propose un classement des isotopies les plus « importantes » des documents du groupe dans le cas d'une cartographie réalisée à partir de dispositifs LUCIA. Ces cartes permettent d'avoir un regard sur les principaux sujets abordés ainsi que sur la répartition des thèmes dans les textes du corpus ;
- des cartes en 2 dimensions animées mettant en évidence l'évolution des thèmes abordés dans les textes lorsque les documents qui constituent le corpus sont datés (c'est par exemple le cas d'un corpus de dépêches d'agence de presse). Ce type de cartes permet

de mettre en évidence les différentes thématiques abordées sur certaines périodes ainsi que leur enchaînement.

Expériences sur corpus réalisées avec ProxiDocs

Dans le but de caractériser la part de la dimension globale intertextuelle dans l'accès aux contenus de documents, nous relatons ici les principes et les résultats de quatre expérimentations logicielles avec la plate-forme ProxiDocs :

- la première sur corpus généraliste avec des ressources elles aussi assez généralistes sous forme de listes de thèmes créées par nos soins ;
- la deuxième sur corpus très spécialisé avec des ressources très spécifiques développées sous formes de dispositifs LUCIA (représentations terminologiques différentielles) également créés par nous même. ;
- la troisième sur un flux documentaire avec des ressources créées par nous même et visant les différentes thématiques abordées à différents moments du flux ;
- la quatrième sur un corpus et des ressources lexicales utilisées dans une recherche en cours sur la terminologie médicale par une chercheuse de l'université de Rouen.

La première expérience réalisée consiste à cartographier un corpus thématiquement hétérogène constitué d'environ 800 articles issus du journal *Le Monde* de 1987 à 1989. Cette expérience (détaillée dans Roy, 2005) prend place dans le domaine de la veille d'informations : elle a pour objectif de découvrir les principaux sujets abordés dans cet ensemble d'articles. Les cartes obtenues à l'issue de cette expérience (présentées en figure 2) ont été réalisées avec un ensemble de 18 thèmes généralistes que nous avons construits, tels la justice, la télévision, l'éducation, etc. La carte des articles met en évidence un nombre très important d'articles de thèmes majoritaires *Politique et élection* (quadrant inférieur droit de la carte) et *Économie et commerce* (quadrant inférieur gauche). Ces observations sont confirmées par la carte des groupes d'articles, la couleur, la taille et la disposition des groupes sur cette carte donnent une information sur les thèmes abordés dans les textes du corpus ainsi que sur leur répartition. En visualisant les rapports des groupes, l'utilisateur peut avoir une idée plus précise des thèmes abordés dans les articles de chaque groupe : il est ainsi facilement observable que le groupe de thème majoritaire Politique et élection contient principalement des articles traitant des futures élections européennes. Les différentes cartes construites durant cette expérience nous ont alors permis de mettre rapidement en évidence les grandes tendances du corpus (principaux sujets abordés dans les textes et groupes de textes abordant des thèmes proches), ce qui est un premier résultat satisfaisant pour une interface de lecture rapide particulièrement utile dans une tâche de veille d'informations.

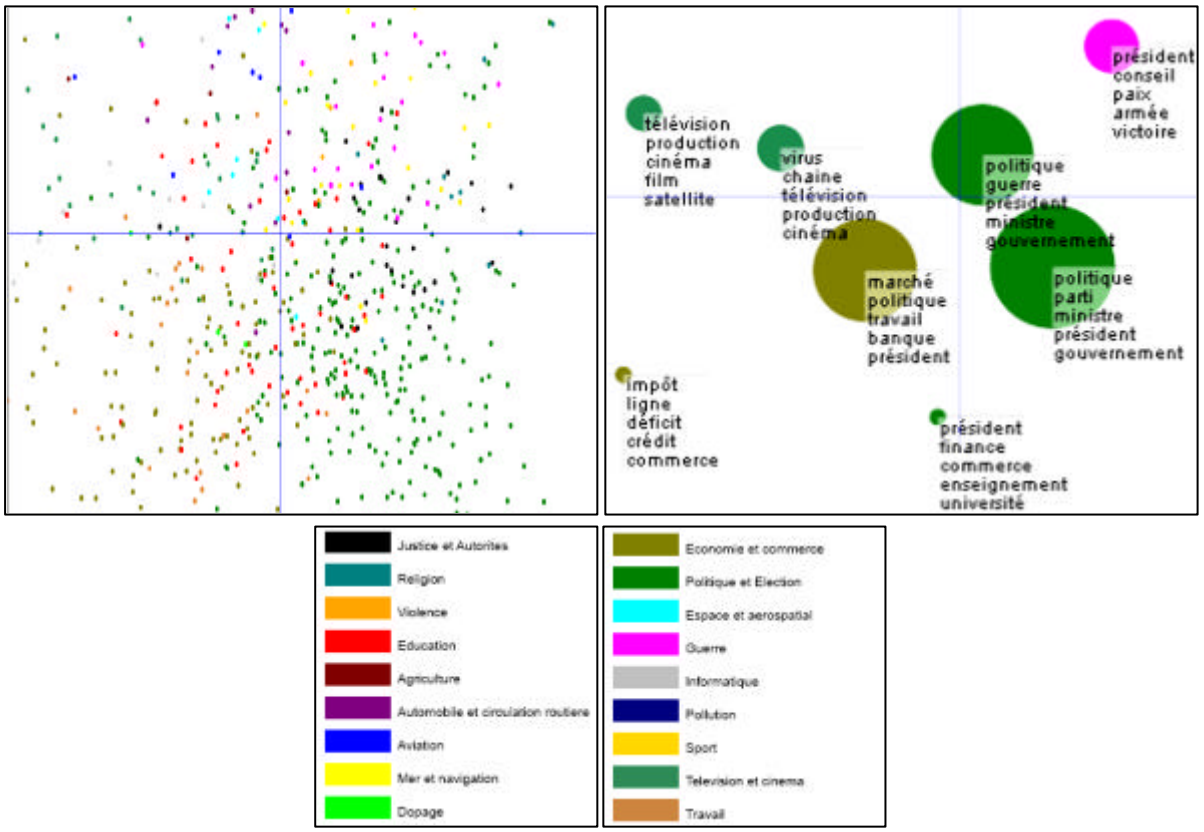

Figure 2 : Cartes thématiques obtenues à partir d'un corpus d'articles de journal et de thèmes généralistes. La carte mettant en évidence les articles du corpus est située en haut et à gauche de la figure, la carte mettant en évidence des groupes de textes est située en haut et à droite de la figure, la légende de couleur de la carte est indiquée sur la partie inférieure de la figure.

La deuxième expérience présentée ici (et détaillée dans (Roy & al., 2005)) consiste à observer un fait de langue : la métaphore conceptuelle au sens de Lakoff et Johnson (Lakoff & al., 1980). Cette observation est faite sur un corpus d'environ 300 articles boursiers issues du journal *Le Monde* entre 1987 et 1989. Les analyses ont porté sur trois métaphores conceptuelles : la *météorologie boursière*, la *guerre économique* et la *santé financière*, un nombre important de ces trois métaphores ayant été observé dans notre corpus d'étude[54]. Les cartes obtenues à l'issue de cette expérience (présentées en figure 3) ont été réalisées avec un ensemble de 3 dispositifs LUCIA représentant les domaines de la météo, la santé et la guerre. Les cartes obtenues nous ont notamment montré une proximité entre des documents contenant des emplois métaphoriques d'un même lexique. Il a aussi été possible d'observer un lien entre le type d'articles (bilans, dépêches, etc.) et les métaphores conceptuelles qui y apparaissaient. De cette manière, nous avons pu déterminer que les métaphores de la *guerre économique* se situaient plutôt dans des dépêches détaillant des événements boursiers ponctuels alors que les

[54] Des extraits de notre corpus d'étude illustrant respectivement ces trois métaphores conceptuelles :
- « Une véritable **tempête** de hausses, alimentée par une marée de capitaux, étrangers pour partie, en quête de placement. » Le Monde 03/08/87
- « Le dénouement dans la **bataille** autour de la première banque commerciale privée du pays a eu peu d'effet sur les cours. » Le Monde 27/02/89
- « La pente fut longue à remonter, et il fallut bien douze mois pour **panser** les **plaies** du sinistre et à commencer à croire à de nouveaux records d'altitude pour le CAC. » Le Monde 01/08/89

métaphores de la *météorologie boursière* et de la *santé boursière* se retrouvaient de manière simultanées dans les bilans boursiers hebdomadaires et mensuels.

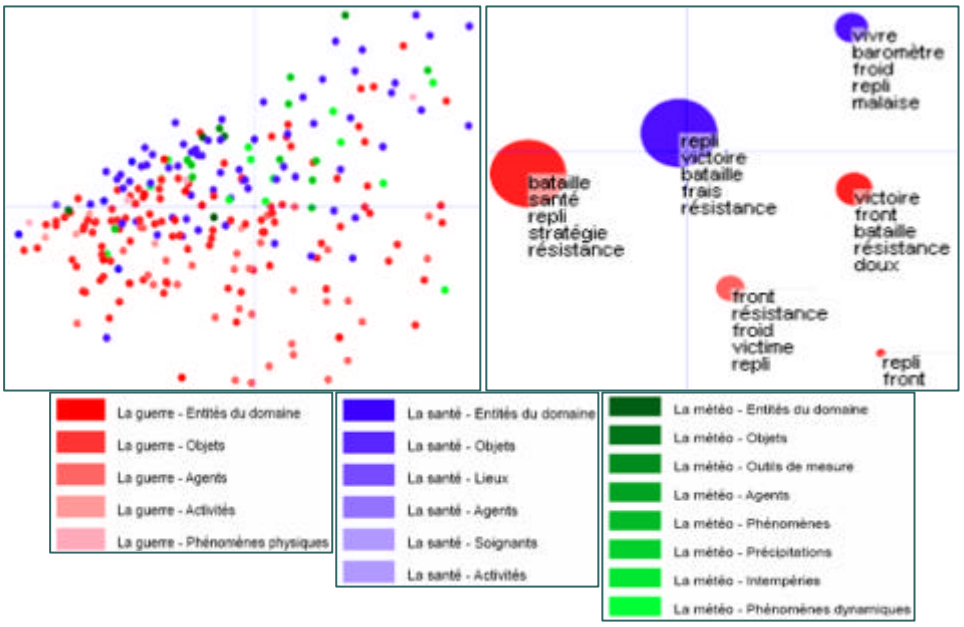

Figure 3 : Cartes thématiques obtenues à partir de notre corpus constitué d'articles boursiers et de dispositifs LUCIA représentant les domaines des métaphores conceptuelles étudiées.

La troisième expérience présentée ici consiste à cartographier un forum de discussion spécialisé portant sur l'apprentissage d'un langage de programmation. Le forum étudié est issu de la plate-forme INES[55]. Il permet à des étudiants de DEUST Technicien des Systèmes d'Information et de Communication d'échanger des messages en rapport avec leur formation. Ce forum est constitué d'environ 200 messages échangés entre le 18/02/2003 et le 27/04/2005 par 27 intervenants différents (enseignants et étudiants). Les cartes de ce forum ont été construites à partir d'un ensemble de 5 thèmes spécialisés que nous avons construits, ces thèmes portant sur l'enseignement, son déroulement, la recherche d'informations, etc. De telles cartes (présentées en figure 4) mettent en évidence les sujets principalement abordés dans les messages. Il est ainsi possible d'observer que la thématique liée au déroulement des enseignements est très majoritaire dans le forum, alors que la thématique liée au contenu même des enseignements est très peu abordée. Ces résultats globaux peuvent alors être des signes pour les enseignants de certaines attentes de leurs étudiants.

[55] http://www.dep.u-picardie.fr/ (consultée le 29-06-2006).

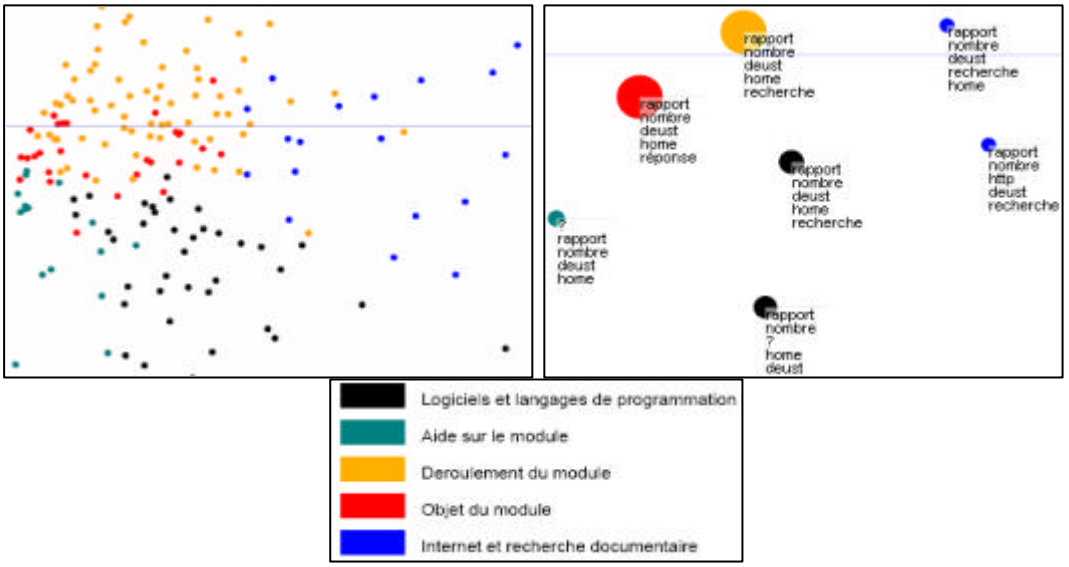

Figure 4: Cartes thématiques obtenues à partir de notre forum de discussion et de thèmes spécialisés.

La quatrième et dernière expérience présentée ici consiste à cartographier un corpus d'articles scientifiques médicaux à partir de ressources lexicales portant sur ce domaine. Cette expérience est en cours de réalisation et s'inscrit dans une recherche sur la terminologie médicale menée par une chercheuse en TAL de l'université de Rouen (Aurélie Néveol (2005)). La carte présentée en figure 5 illustre les premiers résultats obtenus. Elle met ainsi en évidence des sous-ensembles d'articles et caractérise chacun de ces sous-ensembles par ses « métatermes[56] » les plus occurrents. Lors du passage de la souris sur l'un de ces métatermes, les métatermes identiques caractérisant les autres groupes se mettent en relief permettant ainsi de mettre en évidence les métatermes partagés ou non entre sous-ensembles de documents. Cette mise en relief peut alors être particulièrement utile dans une tâche d'indexation de ces groupes d'articles. L'intérêt ici est que l'indexation peut être guidée principalement par la dimension intertextuelle des documents plutôt qu'uniquement par leur contenu.

[56] Les métatermes sont des regroupements de mots-clés réalisés en fonction de spécialités médicales, par exemple le métaterme « *ophtalmologie* » regroupe notamment les termes « *œil* », « *myopie* », « *hypermétropie* » (Soualmia & al., 2002).

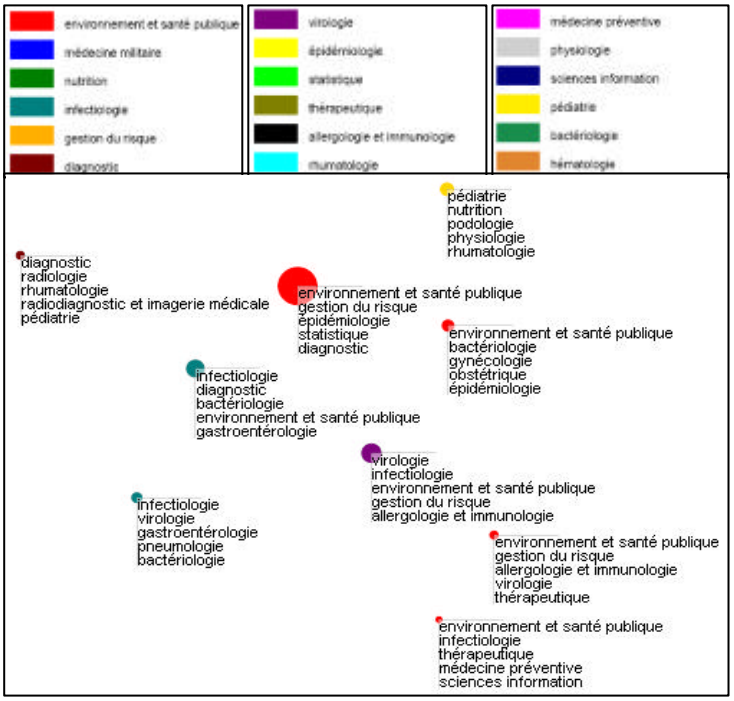

Figure 5: Carte thématique obtenue à partir d'un corpus d'articles médicaux et de ressources terminologiques spécialisées dans ce même domaine.

Perspectives de recherche

Les expériences présentées ci-dessus constituent des premières tentatives de prise en compte de la dimension globale d'un corpus ou d'un flux documentaire dans la contextualisation du sens pour les méthodes d'accès au contenu des documents. Il est clair qu'il faut encore chercher à aller plus loin dans la modélisation de ces méthodes d'accès. L'objectif est la mise au point de modèles opératoires (utilisant des ressources légères) pour les opérations interprétatives et plus particulièrement pour l'afférence. Pour cela il faut continuer à étudier finement des corpus afin d'en dégager différentes formes contextuelles de distribution sémique.

C'est en partie le travail que nous menons dans le cadre du projet ISOMETA visant la caractérisation d'emplois métaphoriques dans des corpus d'actualité boursière. Dans (Perlerin & al. 2005), auquel nous renvoyons le lecteur pour de plus amples détails, nous avons donné quelques descriptions de telles formes de dynamique sémique contextuelle. Par exemple, dans l'extrait de corpus suivant

> Extrait de l'article n°126
>
> *Le Dow Jones par exemple, le thermomètre de la Bourse de New York, qui avait chuté de 508 points (…).*

la lexie *thermomètre* est source d'une métaphore *in præsentia* dont la cible est exprimée par la lexie *Dow Jones*. Entre ces 2 deux lexies, nous avons montré une isotopie d'un sème indiquant la fonction d'un objet qui sert à l'étude et l'analyse (au même titre que d'autres lexies définies dans nos ressources, *graphique* ou *courbe*, partageant également ce sème). Une telle interprétation rend compte de la nature analogique du lien métaphorique. Une autre forme de dynamique sémique nous a permis de nous prononcer sur une autre interprétation d'une métaphore dans l'exemple suivant :

Extrait de l'article n°153
Ce krach était dû (…) à la chute vertigineuse et incontrôlée du dollar, signe que la tempête affecte dorénavant les marchés financiers.

Dans cet exemple, la lexie *tempête* est source d'une métaphore *in absentia* dont la cible n'est donc pas dans le cotexte. Entre *tempête* et *krach* nous avons montré une isotopie indiquant quelque chose évalué comme *mauvais*. De plus le contenu sémique de *tempête* porte un sème potentiellement partageable par plusieurs domaines thématiques qui indique un phénomène *violent*. Cette nature partageable du sème permet de l'actualiser dans le cotexte. Nous en avons déduit cette fois une caractérisation de l'aspect créatif du lien métaphorique.

D'autres pistes de recherche peuvent aussi être explorées pour tenter de rendre compte de mécanismes d'afférence sans pour autant nécessiter le recours à de vastes ressources. Notamment, nous cherchons à l'heure actuelle à exploiter l'importance relative d'une isotopie par rapport à une autre au sein d'un document. L'idée principale est de « positionner » les textes et les groupes de textes (déterminés automatiquement par la plate-forme ProxiDocs par exemple) par rapport aux éléments de plus haut niveau les englobant (respectivement, groupes de textes et corpus) en tenant compte de domaines représentés par l'utilisateur selon le modèle LUCIA. Ce positionnement d'une entité textuelle par rapport à une entité textuelle plus globale la contenant permet d'obtenir des informations pertinentes sur la répartition et la localisation des domaines représentés en corpus. Pour aller plus loin dans de telles analyses, nous tenons compte des particularités du niveau global (que l'on pourrait appeler « signaux forts ») à un niveau plus local (où prennent place ce que l'on pourrait appeler des « signaux faibles »). Pour mettre en œuvre cette prise en considération du niveau global, nous pondérons les classements des isotopies que nous retrouvons d'un document à l'autre selon les deux critères suivants :

- si dans le groupe et dans le corpus, une même isotopie est très présente, alors on diminue son importance dans le groupe (atténuation du signal fort) ;
- si au contraire, dans le groupe, une isotopie est présente et qu'elle l'est moins dans le corpus, alors on augmente son importance (amplification du signal faible).

Ces deux conditions permettent de faire ressortir des propriétés des groupes masquées par les propriétés globales du corpus dont chaque texte hérite. Ce positionnement des groupes par rapport aux corpus aide ainsi à identifier comment les ressources sont projetées sur les différents paliers textuels (texte et groupe) du corpus analysé et en quoi elles permettent de différencier et d'isoler des groupes et des textes.

Ces quelques heuristiques de prise en compte de la dimension globale sont encore au stade de l'expérimentation au sein de l'outil ProxiDocs. L'enjeu est maintenant de les évaluer. Cette évaluation ne pourra être limitée à une quantification technique des résultats produits car, en tant qu'outil individu-centré, ProxiDocs doit faire l'objet d'évaluations extrinsèques au sens de (Spark Jones & al., 1995), c'est-à-dire des évaluations construites sur le recueil et l'analyse des avis des utilisateurs. Il s'agira donc d'une certaine façon de caractériser si l'apport de la dimension globale d'un corpus sur l'accès au contenu des documents est consensuel ou au contraire fait l'objet de variations interpersonnelles.

Conclusion

Plus que jamais, notre objectif reste de poursuivre nos travaux sur la cartographie personnalisée de corpus pour aller toujours plus loin dans la modélisation et l'amélioration des Interactions Homme-Machine (IHM) où un rapport sémiotique au langage et aux textes est central. Dans ce genre d'IHM, nous préférons de loin l'idée d'une instrumentation du sens à celle de la construction du sens. D'une manière imagée, il nous semble qu'un traitement

sémantique informatisé a plus de points communs avec un outil tel un microscope par exemple, c'est-à-dire quelque chose qui nous montre ce qui est déjà là mais qu'on ne voyait pas de cette façon, qu'avec un outil tel un transformateur électrique qui produirait à partir de quelque chose une autre chose qui n'existait pas avant. Il s'agit donc de considérer que le sens tel que le produit une interprétation humaine n'est pas à la portée d'un seul traitement informatique. Cela ne veut pas dire pour autant que les machines ne puissent pas interpréter des textes. Elles le font à leur manière comme les sujets humains le font aussi à leur manière. Nous opposons ici l'idée d'une Interprétation Calculatoire (IC) à celle d'une Interprétation Humaine (IH). Ces deux formes d'interprétation ne sont pas en concurrence car l'IC n'a en aucun cas le but de supplanter l'IH. Au contraire, nous les pensons comme complémentaires dans le sens où une interprétation calculatoire a pour objectif de produire dans l'interaction des traces qui vont participer aux interprétations humaines du ou des utilisateurs.

Comme le calcul rapide des opérations numériques complexes ou encore le traitement immédiat de vastes corpus sont inaccessibles aux capacités cognitives humaines, les finesses de sens (par exemple celles que l'on trouve dans un bon nombre de formes d'humour ou d'interprétations littéraires) ainsi que les mises en relation diverses et variées bien spécifiques à l'interprétation humaine ne sont pas, pour la majeure partie, modélisables dans le cadre d'une interprétation calculatoire. La recherche de ce qui est à la limite des deux formes d'interprétation et qui peut se prêter à une modélisation dans le cadre d'une interprétation calculatoire suscite à l'évidence bien plus de questions que de réponses. La question des rapports entre le global et le local dans la contextualisation du sens nous paraît notamment se situer exactement sur cette limite. L'amélioration des compétences sémiotiques et interactionnelles des machines constitue donc en tout cas un domaine de recherche à part entière au croisement de plusieurs disciplines et pas uniquement un domaine d'ingénierie. En cela, à l'encontre de Rastier (Rastier, 2005, p. 41), nous militons pour une scientificité propre des traitements sémantiques et plus largement des traitements automatiques des langues.

Bibliographie

ASHER N., 1993, *Reference to Abstract Objects in Discourse*, Dordrecht, Kluwer.

BERNERS-LEE T., 1998, *What the Semantic Web can represent ?*, W3C, http://www.w3.org/designissues/rdfnot.html (consultée le 29-06-2006), MANN, W.C., & THOMPSON, S.A.

BEUST P., 1998, *Contribution à un modèle interactionniste du sens*, Thèse de doctorat en Informatique, Université de Caen Basse-Normandie.

BEUST P., 2002, «Un outil de coloriage de corpus pour la représentation de thèmes », dans 6èmes *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*.

BOURIGAULT D., AUSSENAC-GILLES N., 2003, « Construction d'ontologies à partir de textes », dans *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Tome 2, pp. 27-47.

CELEX, 1998, http://www.ldc.upenn.edu/readme_files/celex.readme.html (consultée le 29-06-2006), UPenns, Eds., Actes de *Consortium for Lexical Resources*.

CHARLET, J., LAUBLET, P., REYNAUD, C., 2003, *Web Sémantique*. Rapport de l'Action Spécifique 32 CNRS / STIC. V3. http://rtp-doc.enssib.fr/IMG/pdf/ASWebSemantique2003.pdf (consultée le 29-06-2006).

CLAVEAU V., 2003, *Acquisition automatique de lexiques sémantiques pour la recherche d'information*, Thèse de doctorat en Informatique, Université de Rennes 1.

CONDAMINES A. (dir.), 2005, *Sémantique et corpus*, Hermès, Paris.

COURSIL J., 2000, *La fonction muette du langage*, Ibis Rouge Editions, Petit-Bourg (Guadeloupe).

JACQUEMIN C., BUSH C., 2000, «Fouille du Web pour la collecte d'entités nommées », dans *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Lausanne 187-196.

KAMP H., 1981, « A theory of truth and semantics representation », dans *Formal Methods in the Study of Language* sous la direction de GROENENDIJK, JANSEN & STOKHOF, Amsterdam, Mathematical Centre Tracts.

LAKOFF G., JOHNSHON M., 1980, *Metaphors we live by*, University of Chicago Press, Chicago, U.S.A.

LAVENUS K., LAPALME G., 2002, «Évaluation des systèmes de question réponse », dans *Traitement Automatique des Langues*, vol. 43, n°3/2002, p. 181-208.

NEVEOL A., 2005, *Automatisation des tâches documentaires dans un catalogue de santé en ligne*, Thèse de doctorat en Informatique, INSA de Rouen.

NICOLLE A., 2005, « Réflexivité et auto-référence dans les systèmes complexes », *in Comparaison entre les comportements réflexifs du langage humain et la réflexivité des langages informatiques*, Actes des 12e journées de Rochebrune, pp. 137-148.

PERLERIN V., 2002, « Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes », *in* Actes de *TALN-RECITAL* dans le cadre des *Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Tome 1, pp. 507-516.

PERLERIN V., 2004, *Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse de doctorat en Informatique, Université de Caen Basse-Normandie.

PERLERIN, V., BEUST, P., FERRARI, S., 2005, « Métaphores et dynamique sémique », dans *La Linguistique de Corpus*, sous la direction de G. WILLIAMS, Rennes, Presses universitaires de Rennes, pp. 323-336.

RASTIER F., 1987, *Sémantique interprétative*, Paris, Presses Universitaires de France.

RASTIER F., 1998, *Le problème épistémologique du contexte et le problème de l'interprétation dans les sciences du langage*, *Langages*, n°129, pp.97-111.

RASTIER F., 2001, *Arts et Sciences du texte*, Paris, Presses Universitaires de France.

RASTIER F., 2005, « Enjeux épistémologiques de la linguistique de corpus », dans *La Linguistique de Corpus*, sous la direction de G. WILLIAMS, Rennes, Presses universitaires de Rennes, pp. 31-45.

RASTIER F., CAVAZZA M., ABEILLE A., 1994, *Sémantique pour l'Analyse*, Paris, Masson.

ROY T., 2005, « Une plate-forme logicielle dédiée à la cartographie thématique de corpus », dans *Actes de TALN-RECITAL 2005* dans le cadre des *Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pp. 545-554.

ROY T., BEUST P., 2004, « ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus », dans *Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, pp. 978-987.

ROY T., FERRARI S., BEUST P., 2005, «Cartographie thématique des corpus pour l'étude des métaphores », dans *Actes des Journées de Linguistique de Corpus (JLC)*, WILLIAMS G. Ed., à paraître.

SOUALMIA L.F., BARRY-GREBOVAL C. ABDULRAB H & DARMONI S.J., 2002, « Modélisation et représentation des connaissances dans un catalogue de santé », dans Actes de *Ingénierie des Connaissances (IC)*, pp. 139-149.

SPARK-JONES K. & GALLIERS J. R., 1995, « Evaluating Natural Language Processing Systems: An Analysis and Review ». Number 1083 dans *Lecture Notes in Artificial Intelligence*, Springer.

THLIVITIS T., 1998, *Sémantique Interprétative Intertextuelle*. Thèse de doctorat en Informatique, Université de Rennes I.

VALETTE M. & GRABAR N., 2004, « Caractérisation de textes à contenus idéologiques : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP », dans *Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, pp 1107-1117.

LA CONCEPTUALISATION MÉTAPHORIQUE EN BIOMÉDECINE : INDICES DE CONCEPTUALISATION ET RÉSEAUX LEXICAUX

Sylvie Vandaele, Sylvie Boudreau, Leslie Lubin, Elizabeth Marshman

Université de Montréal, Département de linguistique et de traduction

« Analogie n'est pas identité : les cellules, évidemment, ne parlent pas, au sens où nous autres, êtres humains, doués de langage, nous parlons. » (Lecourt, In : Kordon, 1991 : 9)

De nombreux travaux ont appuyé l'importance de la conceptualisation métaphorique (CM) non seulement dans « la vie quotidienne » (Lakoff, 1980/2003), mais également en littérature (Lakoff et Turner, 1989), en économie et dans les affaires (Mirowski, 2001 ; Perlerin et coll., 2002 ; Koller, 2004), ainsi qu'en sciences, notamment en biologie et en médecine (van Rijn-van Tongeren, 1997 ; Yu, 1998; Fox Keller, 1999), pour ne citer que quelques auteurs. La compréhension de la CM d'un domaine nous semble constituer un outil cognitif puissant dans le processus de traduction (et de rédaction), bien que les études en traductologie soient plutôt rares et récentes (Tabakowska, 1993 ; Stambuk, 1998 ; Schaeffner 2004 ; Temmerman, 2002 ; Vandaele, 2000, 2003), la problématique étant traditionnellement abordée sous l'angle des théories classiques de la métaphore comme élément déviant ou rhétorique (Newmark, 1981).

La plupart des travaux traitant de la métaphore en sciences s'y intéressent sous l'angle terminologique (Gaudin, 1998 ; Bouveret, 1998 ; Dury, 1999 ; Temmerman, 2000, 2006 ; Oliveira, 2003). Nous nous concentrons, pour notre part, sur les aspects phraséologiques, qui nous paraissent véhiculer une composante essentielle des modes de conceptualisation. Ce choix nous amène à privilégier les unités lexicales prédicatives telles que le verbe, assez souvent laissé pour compte en métaphorologie, bien qu'essentiel (Duvignau, 2002). Notre objectif général est de caractériser finement, sous les angles lexical et cognitif, les différents modes de conceptualisation spécialisés tels que les textes biomédicaux les révèlent, excluant pour le moment ce qui relève de la vulgarisation (voir Duvignau, 2002; Collombat, 2003). Au sein du vaste domaine qu'est la biomédecine, la biologie cellulaire et moléculaire (Vandaele, 2003, 2004, 2005 ; Vandaele et Lubin, 2005) et l'anatomie (Lubin, 2006) sont plus précisément ciblées. Outre leur importance fondamentale, il se trouve que la biologie cellulaire et moléculaire constitue un réservoir extrêmement riche d'expressions témoignant de multiples modes de conceptualisation (Kordon, 1991). Quant à l'anatomie, domaine qui pourrait paraître banal en raison de son objet d'étude (quoi de plus familier que le corps humain?), elle recèle une quantité insoupçonnée de représentations conceptuelles et de variations terminologiques et phraséologiques.

Le cadre théorique se réclame essentiellement de la linguistique cognitive, notamment des travaux de Lakoff sur la CM (Lakoff, 1987/2003; 1993), ainsi que ceux de Talmy (2001) pour les concepts de factivité et de fictivité. Les travaux de Fauconnier et Turner (1998) seront sollicités pour l'intégration conceptuelle (*blending*). L'étude lexicale fait appel à une analyse actantielle (Tesnière, 1965 ; Mel'èuk et coll., 1995) et, lorsque cela est pertinent, aux fonctions lexicales (Mel'èuk et coll., 1995). Bien qu'il s'agisse d'outils inspirés par la logique formelle, nous faisons nôtre la position de Le Ny (1979 : 13-14) qui affirme que « *en aucune occasion il ne peut exister d'objet d'étude sémantique qui ne soit, en définitive, de nature psychologique* », mais croit « *pleinement justifiée l'application à la sémantique d'une formalisation empruntée à la logique* ». Cela « *ne signifie nullement que le parleur (…) fonctionne de façon logique* », mais que le chercheur « *essaie de fonctionner de façon logique, c'est-à-dire conformément à des règles qu'il se donne* ».

Enfin, nous insistons sur le fait que l'objet de nos recherches porte, avant tout, sur les *modes de conceptualisation* plutôt que sur la métaphore au sens large : bien que le mot *métaphore* ait vu son sens revisité par Lakoff (projection d'un cadre cognitif source sur un cadre cognitif cible au plan de la pensée), trop souvent il véhicule encore le sens qui lui est le plus souvent attribué, celui qui est consigné dans les dictionnaires de langue générale, à savoir une sorte de comparaison (Duvignau, 2002 : 30), ou encore, selon l'école de pensée : déviance, figure de style, procédé rhétorique. L'objet d'étude n'est en aucun cas en rapport avec une quelconque déviance, c'est la manière courante de conceptualiser le monde biologique qui est ici envisagée. De plus, aborder la question sous l'angle du « mode de conceptualisation » ouvre la porte, ultimement, à des modes autres que métaphoriques, par exemple la conceptualisation métonymique.

1. Méthodologie : identification et repérage des ICM en corpus

1.1. Corpus comparables en anglais et en français

Comme Lakoff, nous analysons la CM à partir d'expressions linguistiques, et comme d'autres chercheurs (Perlerin et coll., 2002 ; Charteris-Black, 2004 ; Koller, 2004 ; Deignan, 2005), nous faisons appel à des corpus. Pour la biologie cellulaire, qui est plus spécifiquement abordée dans le présent article, nous avons construit deux corpus comparables (c'est-à-dire non traduits) constitués de textes *spécialisés* traitant de biologie cellulaire et moléculaire en anglais et en français, totalisant environ 300 000 et 500 000 mots respectivement (voir Vandaele, 2005 pour le détail de leur constitution). Les exemples en anatomie, destinées à étoffer notre propos, sont tirés du travail de Lubin (2006), qui visait à analyser les formes verbales utilisées en anatomie pour décrire le positionnement des artères, des veines, des muscles et des nerfs. Pour ce faire, deux corpus, également en anglais et en français, ont été constitués par des extraits d'ouvrages spécialisés d'anatomie descriptive retenus notamment pour leur statut de référence incontournable.

1.2. Identification des indices de conceptualisation métaphorique

La difficulté de l'identification des expressions linguistiques métaphoriques est que la CM est, par essence, un phénomène cognitif. Par conséquent, aucune approche formelle ne peut être envisagée. Aucune caractéristique syntaxique ne peut permettre d'identifier les expressions métaphoriques (Tamba, 1981 ; Tamine, 1978 ; Duvignau, 2002), bien que les différentes catégories syntaxiques puissent être concernées (ce que Duvignau (2002 : 36) appelle

« l'éclatement syntaxique de la métaphore »). Par conséquent, on ne peut avoir recours qu'à ce que Deignan nomme « *informed intuition* » (2005 : 93), qui correspond, pour nous, aux « connaissances linguistiques et extralinguistiques du locuteur assistées par des données ».

La polymorphie du phénomène métaphorique appellerait une discussion détaillée des critères d'identification (voir par ex. Eco, 1988/2006 : 139-189). Pour résumer, l'identification des ICM opérant en fonction de critères cognitifs, il n'est pas pertinent de « plaquer » les différents modèles appliqués à la métaphore (substitutifs, interactionnels, comparatifs, analogiques), qui d'ailleurs privilégient généralement la forme canonique de la métaphore nominale (Duvignau, 2002). L'objet d'étude étant les modes de conceptualisation, nous avons adopté une stratégie très proche de celle décrite par Talmy (2001). Le critère d'identification d'une expression métaphorique pertinente est la perception d'une « dissonance cognitive » par le sujet, laquelle émerge lorsque celui-ci constate que le référent dont traite le discours peut être conceptualisé de deux manières simultanées, ce que Talmy (2001 : 101, 135-137) a décrit sous le nom de « représentation fictive » (la moins véridique), et de « représentation factive » (la plus véridique). Les deux représentations « s'opposent », elles sont donc dissonnantes. L'élément lexical générant cette impression a été baptisé « indice de conceptualisation métaphorique » (ICM ; Vandaele et Lubin, 2005).

Ainsi, dans le titre de paragraphe présenté dans l'exemple (1) ci-dessous, l'ICM *passage* évoque une représentation mentale de déplacement, qui peut susciter l'impression que les récepteurs se déplacent sept fois en passant à travers la membrane.

(1) « Structure des récepteurs à sept passages membranaires » (Étienne, 1999 : 180)

Or, il n'y a aucun déplacement[57], ainsi qu'en témoignent l'explication fournie dans le paragraphe qui suit[58] ou une illustration[59]. Selon Talmy, la représentation de déplacement fictif correspond au sens de *passage*, tandis que la représentation factive correspond à ce que nous savons de la situation décrite (le récepteur ne se déplace pas). Nous modulons cette interprétation en décrivant deux sens pour *passage*, une des lexies correspondant à ce que nous avons appelé « la lexie source » (celle qui dénote un déplacement), l'autre, «la lexie cible » (celle qui apparaît en contexte et qui dénote non pas un déplacement, mais un positionnement spatial, en l'occurrence la façon dont la chaîne protéique linéaire constituant le récepteur est disposée au sein de la membrane). Cette modulation nous a ainsi permis de proposer dès 2003 que la conceptualisation métaphorique procédait d'une projection de la structure actantielle de la lexie source sur celle de la lexie cible (Vandaele, 2004, 2005). Cette hypothèse a été adoptée par d'autres en vue d'une application à la génétique en espagnol (Vidal et Cabré, 2006). De manière intéressante, nous avons constaté depuis que Eco (1988/2006 : 173) avait déjà proposé une approche semblable. Il est par ailleurs certain que pour les unités terminologiques n'ayant pas d'actant sémantique (comme *cellule*, dénomination dont l'analyse nécessite d'ailleurs une analyse diachronique (Dury, 1999), une analyse par traits sémantiques sera plus pertinente (Perlerin et coll., 2002). Les deux approches sont clairement complémentaires (Le Ny, 2005 : 301-346). Soulignons que l'étude des indices de conceptualisation métaphorique prédicatifs, par la méthode que nous avons adoptée, se fait nécessairement en synchronie.

---

[57] Talmy (2001 : 99) illustre le déplacement fictif notamment avec : « *The fence goes from the plateau to the valley* ».

[58] *« La structure moléculaire de ces récepteurs est déduite de différents types de travaux (physiques, chimiques, pharmacologiques, etc.). Ils comprennent (…):*
- un domaine transmembranaire, constitué de 7 segments transmembranaires hydrophobes comprenant environ 20 à 25 aa [acides aminés], formant 7 hélices alpha. Ces segments sont reliés les uns aux autres par 6 boucles hydrophiles, dont 3 sont intracellulaires et 3 extracellulaires. » (Étienne, 1999 : 187)

[59] Voir, par ex., http://www.cnsforum.com/imagebank/item/D_struc_level2/default.aspx

Le déplacement fictif constitue l'une des représentations fictives les plus présentes en anatomie, où les vaisseaux (artères, veines) et les nerfs sont couramment conceptualisés comme des entités mobiles suivant un « parcours » :

(3) « L'artère méningée moyenne, volumineuse, ***monte*** *verticalement en dedans du ptérygoï dien externe,* ***traverse*** *une boucle formée par le nerf auriculo-temporal et* ***pénètre*** *dans le crâne par le trou petit rond. »* (Rouvière, 1991 : 208)

(4) « The medial supraclavicular nerves ***run*** *inferomedially across the external jugular vein (…). »* (Gray, 1989 : 1128)

Dans ces différents exemples, *monte*, *traverse*, *pénètre* en français, et *run* en anglais constituent les ICM induisant une représentation mentale fictive. Ce type de conceptualisation, qui correspond à une représentation visuelle impliquant un déplacement imaginaire, est assez facile à identifier. Certains modes de conceptualisation sont moins directs, car ils ne font pas intervenir la perception, mais plutôt des connaissances extralinguistiques moins immédiates. Ainsi, dans l'exemple (5), l'identification de l'ICM *communauté* impose de savoir que ce sont des êtres vivants qui forment, de façon prototypique, une communauté, plus précisément des êtres humains, et non pas des cellules.

(5) « Dans un organisme, les cellules forment une ***communauté*** *au sein de laquelle les échanges sont permanents. »* (Alfandari, 1999 :1148)

Le fait que l'accent soit délibérément mis sur la question de la conceptualisation entraîne certaines conséquences. En premier lieu, l'identification d'un ICM implique un certain degré de saillance cognitive des représentations prototypiques qui lui sont associées, ce qui est plus facilement accessible non seulement aux locuteurs natifs qu'aux non-natifs, mais aussi à ceux qui connaissent le domaine de spécialité envisagé. Il sera parfois nécessaire « d'assister "l'intuition" » par une recherche complémentaire faisant appel à des sources externes au corpus étudié (autres corpus de différents domaines, dictionnaires, etc.). La systématisation de telles recherches, pour réduire le plus possible le caractère subjectif de l'analyse, est à établir. Un consensus entre différentes personnes travaillant sur le même corpus ainsi qu'un travail de révision systématique (Perlerin et coll., 2002) est pour le moment le meilleur moyen de garantir la reproductibilité des résultats, la variabilité des représentations cognitives interindividuelles constituant le principal obstacle à ce type d'approches (Talmy, 2001). Cependant, l'existence de la conceptualisation métaphorique n'étant plus à démontrer, il nous paraît essentiel d'avancer, bien qu'avec prudence, vers un « démontage » de son mécanisme.

Par ailleurs, la question de la lexicalisation n'est pas envisagée en tant que critère d'identification. Bien entendu, l'extension de sens métaphorique est l'un des mécanismes de la polysémie, mais ce qui nous intéresse ici, ce n'est pas de savoir si la lexie cible est lexicalisée ou non. De fait, nous avons été amenées à étendre l'hypothèse, initialement restreinte aux composantes phraséologiques, aux dénominations métaphoriques prédicatives (mais nous n'avons pas traité ce problème dans le présent article). Parfaitement lexicalisées en biologie, comme *canal*, *transporteur*, ce sont des indices de conceptualisation dont il importera de décrire le comportement sémantique, ainsi que la cohérence avec les autres indices présents dans le corpus. Quoiqu'il en soit, si la lexie cible est fréquemment utilisée (et contrairement à l'idée reçue, les acceptions métaphoriques sont souvent plus fréquentes en corpus que les acceptions dites « premières » (Deignan, 2005 : 94), il est probable qu'elle est ou qu'elle sera lexicalisée – d'où l'intérêt de travailler à partir de corpus afin d'appréhender le paramètre de l'usage. La question de la lexicalisation est d'autant moins pertinente que la représentation phraséologique des langues de spécialité est encore peu développée. Enfin, puisque la vectorialité de la projection

lexie source - lexie cible découle de la coexistence des deux représentations, fictive et factive, elle n'a rien à voir, *a priori*, avec un ordre des acceptions dans une entrée de dictionnaire – bien qu'il puisse y avoir correspondance.

1.3. Repérage des indices de conceptualisation métaphoriques dans le corpus

Le repérage des ICM consiste à annoter les textes convertis en format électronique à l'aide du langage XML. La structure informatique et les balises utilisées sont décrites ailleurs (Lubin, 2006 ; Vandaele et Boudreau, 2006) et l'annotation elle-même fera ultérieurement l'objet d'un guide détaillé. Brièvement, pour le corpus de biologie cellulaire, trois balises (ou éléments) ont été utilisées : `<concInd>`, pour annoter l'indice de conceptualisation, `<lingEl>`, pour les réalisations des actants dans la phrase, `<col>`, pour repérer les collocatifs des ICM et les caractériser à l'aide de fonctions lexicales. Chacune des balises contient des attributs : `lem`, pour indiquer la forme lemmatisée, `id`, qui confère à l'élément un numéro arbitraire mais unique dans la phrase, act_n, pour pointer vers les réalisations des actants de l'ICM dans la phrase et met_n, pour caractériser la projection métaphorique. Lorsque `<col>` s'applique, `flRef` pointe vers le mot-clé de la fonction lexicale, `fl` indique le nom de la fonction, et `val`, sa valeur sous forme lemmatisée.

(6) «`<phr #>` *Dans un organisme, les* `<lingEl id="3" lem="cellule">`*cellules* `</lingEl> <col fl="IncepOper1" flRef="1" id="2" val="former [ART] communauté" lem="former">`*forment*`</col>` *une* `<concInd flRef="3" met1="personne" fl="Mult" id="1" act1="3">`***communauté***`</concInd>` *[…].*`</phr>` *» (Alfandari, 1999 : 1148)*

Les balises utilisées pour le corpus d'anatomie sont sensiblement les mêmes, hormis certaines particularités liées au projet.

À l'aide d'un formulaire d'interrogation, il est alors possible d'extraire les données voulues de façon à obtenir des données quantitatives ou qualitatives à partir des corpus. Le présent article se concentre sur des données semi-quantitatives en français, et sur l'analyse de la structure sémantique des indices de conceptualisation.

2. Caractéristiques des indices de conceptualisation métaphorique

2.1. Catégories lexicales concernées

Bien que le volume de texte annoté (16 683 mots en français, 12 146 mots en anglais) soit relativement modeste en regard de la totalité du corpus (environ 3,5 %), les données recueillies nous paraissent être représentatives des phénomènes observables dans les corpus complets en raison du nombre d'occurrences repérées, notamment pour les modes de conceptualisation les plus saillants et la cohérence des réseaux lexicaux. Le fait que différents modes de conceptualisation aient été identifiés permet de penser que les sujets annotateurs n'étaient pas influencés par une conceptualisation particulière. Dans les deux langues, les indices de conceptualisation se répartissent entre les noms, (F : *rôle, territoire, région…* ; A : *communication, family, region…*), les verbes (F : *coloniser, coder… ;* A : *to act, to participate…*) et les adjectifs (F : *responsable, capable… ;* A : *responsible, active…)*. Pour le moment, aucun adverbe n'a été identifié, bien que les deux corpus en contiennent. En français,

les 721 occurrences[60] d'ICM relevées se répartissent en 44 noms (~51 %), 28 verbes (~33 %) et 14 adjectifs (~16 %). La répartition est du même ordre en anglais, bien que les ICM semblent plus nombreux, mais il est bien entendu que des données quantitatives fiables se prêtant à une analyse statistique ne pourront être obtenues que lorsqu'une plus grande fraction des corpus sera annotée et révisée. Il faut évidemment s'attendre à ce que la liste des ICM identifiés s'allonge dans la suite du travail.

De par leur nature, les ICM verbaux et adjectivaux sont des unités lexicales prédicatives (Tableau 1, pour le français). Par ailleurs, nous avons relevé un certain nombre d'ICM nominaux prédicatifs (Tableau 2). Enfin, le nombre d'occurrences de chacun des ICM varie de 1 (pour 32 ICM) à 82 (*expression*). Le tableau 3 présente les ICM les plus fréquemment repérés.

Adjectifs et adj. participiaux	**Verbes**	
actif	agir	intervenir
capable	coder	libérer
compétitif	coloniser	lier
ancré	coopérer	migrer
enchâssé	déplacer	mobiliser
inactif	diriger	mourir
incapable	donner naissance	reconnaître
immature	élucider	recruter
impliqué	exprimer	rencontrer
jeune	fixer	se déposer
programmé	identifier	se fixer
responsable	induire	se lier
sevré	interagir	séquestrer
traduit	interférer	s'exprimer

Tableau 1 - Indices de conceptualisation adjectivaux et verbaux

[60] Les chiffres présentés dans l'article sont obtenus à partir de la partie annotée du corpus, sauf lorsque cela est précisé.

Nom	Structure actantielle	Nom	Structure actantielle
architectonique[61]	~ de X	cible	~ X de Y
ancrage	~ de X dans Y	transcription	~ par X de Y
cascade	~ de X	chaîne	~ de X
résistance	~ de X à Y	expression	~ par X de Y
candidat	~ de X pour Y	identification	~ par X de Y
capacité	~ de X pour Y	langage	~ de X
communauté	~ de X	leurre	~ utilisé par X pour tromper Y
compétition	~ de X à l'égard de Y	liaison	~ de X à Y
expression	~ par X de Y	libération	~ par X de Y
famille	~ de X	machinerie	ensemble de X fonctionnant de façon coordonnée pour un but Y
implication	~ de X dans Y	message	~ de X à Y envoyé par Z au moyen de W
interaction	~ de X avec Y	passage	~ de X dans Y/de Y à Z
intervention	~ de X dans Y	porteur	~ X de Y
messager	~ d'un message X de Y à Z	recrutement	~ par X de Y
parenté	relation entre les membres X d'une famille	repos	~ de X
partenaire	~ X de Y	sevrage	~ de X par Y par rapport à Z
population	~ de X de territoire Y	signal	~ de X envoyé à Y par Z
relais	~ entre X et Y	survie	~ de X
réponse	~ de X à Y	territoire	partie de X occupée par Y
rôle	~ de X en tant que Y dans Z	transmission	~ par X de Y à Z
migration	~ de X de Y à Z	voie	~ de X
mort	~ de X	voisin	~ X de Y

Tableau 2 - Indices de conceptualisation nominaux

ICM	Nbre d'occurrences
expression	82
impliqué	50
famille	47
rôle	39
interaction	31
induire	30
signal	29
responsable	26
réponse	25
interagir	25
identifier	24
voie	20

Tableau 3 – Indices de conceptualisation les plus fréquents

[61] Dans le domaine, anglicisme ayant un sens proche de architecture (« La mise en place de l'architectonique radiaire dépend de Reelin, mais également de Dab1, VLDLR et ApoER2 qui sont exprimés par les cellules de la plaque corticale » (Bar et Goffinet, 1999 : 1284). Le cas des ICM résultant d'interférences linguistiques possibles serait à étudier de près.

Le fait que les ICM nominaux soient à peu près à égalité avec les ICM verbaux et adjectivaux pris ensemble est compatible avec d'autres travaux qui soulignent l'importance des expressions métaphoriques autres que nominales, verbales en particulier (par ex., Duvignau 2002). Par ailleurs, 12 ICM sur 86 totalisent à eux seuls 428 occurrences sur 721, ce qui témoigne du fait que ces unités sont relativement fréquentes.

2.2. Actants sur lesquels opère la conceptualisation métaphorique

Dix-neuf ICM sont monoactantiels, soixante et un sont biactantiels, cinq sont triactantiels et un seul a quatre actants (*message* ; voir Vandaele, 2005 pour une analyse détaillée). L'examen des ICM monoactantiels permet déjà de dégager le mode de conceptualisation prédominant, à savoir que les molécules et les cellules sont conceptualisées, selon le cas, comme des personnes ou des êtres vivants, ce que confirme l'analyse des ICM multiactantiels (données non présentées).

ICM (lemmatisé)	**Réalisations des actants de la lexie cible**	**Classes des actants de la lexie cible[62]**	**Paraphrase de l'ICM en bio. cell. et moléc.**	**Classes des actants de la lexie source**	**Paraphrasage de la lexie source**
actif[63] / *inactif*	*protéine, enzyme, kinase, facteur NF-KB récepteur, précurseur, forme (de molécule), sous-unité complexe*	MOLECULE PARTIE DE MOLECULE, ASSOCIATION DE MOLECULES	qui peut avoir un effet / qui ne peut avoir d'effet	PERSONNE	qui fait une action / qui ne fait pas ou ne peut faire d'action
agir	*facteur de transcription, médicament, protéine*	MOLECULE SUBSTANCE	avoir un effet	PERSONNE	faire une action
chaîne	*a) ~ peptidique* *b) ~ métabolique* *~ respiratoire*	A) MOLECULE (ACIDES AMINES) B) EVENEMENT PHYSIOLOGIQUE	a) ensemble d'acides aminés reliés les uns aux autres linéairement b) suite d'évènements physiologiques	ARTEFACT	objet constitué de maillons
communauté	*cellule*	CELLULE[64]	ensemble de cellules agissant de façon coordonnée	PERSONNE	groupe social partageant certaines caractéristiques

[62] En l'absence de ressource fiable, les classes ont été déterminées de façon *ad hoc*, comme genre prochain pour une définition à l'intérieur du domaine considéré, avec le critère supplémentaire que le nom de classe doit être l'unité la plus générique qui accepte l'ICM identifié comme actant. La problématique des classes remonte à l'antiquité et à la question des arbres de Porphyre (Eco, 1988/2006 : 63-137) et constitue toujours un problème de fond.

[63] On trouve dans le reste du corpus le couple *transport actif/transport passif*, qui correspond à l'idée d'un phénomène réclamant ou non de l'énergie. La conceptualisation est alors différente.

[64] CELLULE est l'exception au deuxième critère explicité en note 8 : il est déjà le plus générique.

ICM (lemmatisé)	Réalisations des actants de la lexie cible	Classes des actants de la lexie cible[62]	Paraphrase de l'ICM en bio. cell. et moléc.	Classes des actants de la lexie source	Paraphrasage de la lexie source
compétitif	*inhibiteur*	MOLECULE	pouvant prendre la place d'un autre ligand sur le récepteur	PRODUIT (?) PERSONNE (?)	qui peut supporter la concurrence (?) qui aime la compétition (?)[65]
famille	*protéines G, lipide-kinases, protéine-kinases, aquaporines, glycoprotéines,* etc.	MOLECULE	ensemble de molécules possédant une structure apparentée	PERSONNE	ensemble de personnes apparentées
immature	*protéine* *neurone* *cervelet**	MOLECULE CELLULE ORGANE	qui n'a pas atteint la maturité fonctionnelle	ETRE VIVANT	qui n'a pas atteint la maturité physiologique ou psychologique
jeune	*cellule*	CELLULE	qui est apparue depuis peu de temps	ETRE VIVANT	peu âgé
langage	*cellule*	CELLULE	fonction de communication au moyen de signaux électriques ou de molécules	HUMAIN	fonction d'expression de la pensée au moyen de signes
mort	*cellule*[66]	CELLULE	arrêt du fonctionnement	ETRE VIVANT	arrêt des fonctions vitales
repos *(au repos)*	*cellule* *membrane cellulaire*	CELLULE PARTIE DE CELLULE	sans activité	HUMAIN/ANIMAL	qui se repose
survie	*cellule* *embryon** *organisme**	CELLULE	fait d'échapper à l'arrêt du fonctionnement	ETRE VIVANT	fait d'échapper à la mort

* autres occurrences relevées dans l'ensemble du corpus

Tableau 4 – Indices de conceptualisation monoactantiels

En ce qui concerne les ICM biactantiels, c'est, selon le cas, le premier actant (7, 8) ou le deuxième (9) qui subit la conceptualisation, ou les deux (10, 11). Toutefois, le schéma le plus courant est celui dans lequel le premier actant est conceptualisé.

- *capable (12 occurrences) : **X** est ~ de faire Y*

*(7) « En se fixant sur leurs récepteurs, certains types de **ligands** sont **capables** de déclencher une action dans la cellule. »* (Étienne, 1999 : 180)

- *rôle (39 occurrences) : ~ de **X** dans Y*

*(8) « La **protéine ADAM1O**, initialement purifiée chez le boeuf pour sa capacité de dégrader la protéine basique de la myéline, joue aussi un **rôle** dans la détermination des cellules neurales. »* (Alfandari, 1999 : 1149)

[65] Cette acception, occasionnelle en français, est empruntée à l'anglais *competitive*. Il est probable que *compétitif*, dans ce domaine de spécialité, résulte lui aussi d'un emprunt à l'anglais. Il se pourrait que la CM soit transférée de l'anglais au français, avec plus ou moins de saillance selon l'usage des lexies empruntées.

[66] Exprimé par un adjectif relationnel : *cellulaire*.

- *élucider (3 occurrences) : **Y** est ~ (par X)*[67]

*(9) « Récemment, le **mécanisme de clivage** du TNF- a été **élucidé**: il implique l'activité protéolytique de la protéine ADAM17 (TACE) [3]. »* (Alfandari, 1999 : 1149)

- *interaction (31 occurrences) : **X** ~ avec **Y***

*(10) « ADAM2 intervient dans **l'interaction spermatozoïde-ovule**, le rôle d'ADAM1 n'est pas encore clairement défini. »* (Alfandari, 1999 : 1149)

- *coopérer (3 occurrences) : **X** ~ avec **Y***

*(11) « Le complexe, suivant la reconnaissance spécifique d'un motif XRE, exerce une transactivation génique durant laquelle **AHR coopère** avec **Sp1** et **Arnt** avec **CBP/p300** et/ou **Sp1**. »* (Lesca, 1999 : 1383)

2.3. Modes de conceptualisation

2.3.1. Conceptualisation des entités biologiques comme des personnes

Le mode de conceptualisation métaphorique le plus général est celui qui attribue aux molécules biologiques et aux éléments cellulaires une volonté, comme des personnes. L'emploi de verbes d'action et de la voix active y contribue, de même qu'un grand nombre d'occurrences d'ICM tels que *rôle* en français et *role* en anglais[68] ou *responsable*[69].

Le cas de *responsable* est particulièrement intéressant, car il transgresse la norme générale du français, et ce probablement sous l'influence de l'anglais. La plupart du temps, le premier actant de *responsable* est exprimé par un terme dénotant une partie de molécule, une molécule, une cellule ou un organisme, c'est-à-dire une entité. Dans ce cas, il permet d'exprimer une fonction :

*(12) « Ces canaux sont tous constitués d'une sous-unité principale **responsable** des transferts ioniques spécifiques. »* (Alliet 1997 : 479)

*(13) «La rhodopsine est la molécule **responsable** de la capture des photons incidents. »* (Alliet 1997 : 490)

Cet usage est habituel en biologie, bien que selon les normes générales de la langue française, il soit perçu comme erroné, *responsable* n'est pas censé s'employer pour les « choses », mais uniquement pour les personnes[70]. *Responsable* provoque ainsi la « dissonance cognitive » qui fait de lui un ICM dans le domaine de la biologie cellulaire et moléculaire, et ce de façon cohérente avec les autres ICM témoignant de la conceptualisation des molécules et des cellules en tant que personnes ou êtres vivants. (Nous discuterons plus loin d'un facteur évidemment crucial, qui est la fréquence avec laquelle un phénomène se produit, ce qui signifie qu'un mode de conceptualisation donné, pour être généralisé dans un domaine, doit être corrélé à un réseau lexical à la fois diversifié et se manifestant fréquemment.)

---

[67] X ('le chercheur') est rarement exprimé, en raison de l'emploi de la voix passive dans les textes scientifiques lié à l'effacement du sujet. C'est pourquoi nous exprimons la structure actancielle de cette façon, ce qui n'est pas canonique.

[68] Environ 850 occurrences de *rôle* dans tout le corpus français; environ 260 occurrences de *role* dans tout le corpus anglais.

[69] Environ 320 occurrences de *responsable* dans l'ensemble du corpus français; une centaine d'occurrences de *responsible* dans l'ensemble du corpus anglais.

[70] « *Ce mot ne se dit que d'une personne; une chose ne peut être la cause d'un fait fâcheux (elle ne peut être responsable)*. La chaussée glissante a causé (*et non* *est responsable) de nombreux accidents. » (De Villers, 1997 : 1267).

Par ailleurs, dans un certain nombre de cas, le premier actant dénote un fait, et *responsable* exprime plutôt un rapport de causalité :

(14) « L'entrée du calcium serait ***responsable*** *d'une potentialisation synaptique pendant une longue période. »* (Alliet 1997 : 476)

Ici aussi, pour certains[14], cet usage s'éloigne de la norme de la langue française. Il est fort possible que l'extension de l'usage de *responsable* se soit opérée sous l'influence de l'anglais, car *responsible* ne subit pas la même restriction[71]. De fait, ce type d'usage semble être de plus en plus fréquent et pourrait même s'installer pour longtemps, peut-être grâce à la conceptualisation métaphorique. Un type d'extension de sens, à la fois sous l'influence d'un mode de conceptualisation particulier et d'une autre langue, avait déjà été décrit pour *être impliqué dans*, qui évoquait la métaphore de l'enquête dans le domaine médical (Vandaele, 2003).

Dans le cas de *rôle* et de *responsable*, la CM opère relativement simplement, par la projection de la classe d'actants de la lexie source sur la classe d'actants de la lexie cible.

2.3.1. Autres modes de conceptualisation

La métaphore du langage et du texte appliquée au fonctionnement des gènes (*transcription, traduction, code, expression*) a déjà été abondamment soulignée (par ex. Temmerman, 2000) et nous avons déjà évoqué celle de la transmission des signaux et des messages (Vandaele 2004, 2005). Certains ICM témoignent de modes de conceptualisation particuliers, mais de façon beaucoup plus dispersée (la liste n'est évidemment pas exhaustive) :

- *machinerie (cellulaire)* : ensemble de X fonctionnant de façon coordonnée pour un but Y. Ici, ce sont les composants de la cellule qui sont conceptualisés comme les éléments constituant une machinerie (X) et qui coopèrent pour faire fonctionner le tout et assurer la « fonction de la cellule » (Y). *Cellulaire* est un adjectif relationnel mis pour un circonstant (et non un actant) de *machinerie*.

- *ancrage de X dans Y* (*ancrage des protéines dans la membrane*) : la conceptualisation évoquée ici est celle du bateau (X) ancré dans le fond de la mer (Y).

- *chaîne d'acides aminés* : les protéines sont conceptualisées comme des chaînes dont les maillons sont constitués par des acides aminés.

Certaines formulations mettent en évidence un déplacement fictif. Nous avons vu, avec l'exemple 1 :

(15) « Structure des récepteurs à sept ***passages*** *membranaires »* (Étienne, 1999 : 180)

La structure actantielle de la lexie cible *passage* est la suivante, X étant exprimé par *récepteur*, et Y par *membranaire* ('dans la membrane') : '~ de X dans Y'. Dans le cas de la lexie source, X est une entité capable de déplacement (*le passage des voitures sur le pont est toujours difficile*). Par conséquent, la coexistence des représentations fictive et factive mobilise les deux structures actantielles, la classe des actants X de la lexie source se projetant sur la classe des actants de la lexie cible. Il n'est cependant pas nécessaire que les structures actantielles des lexies source et cible soient identiques : en fait, l'inverse semble fréquent.

On remarquera que, dans ces différents exemples, les classes d'actants se projetant les unes sur les autres restent dans la catégorie des entités, mais ce n'est pas toujours le cas (voir plus loin). Mais il est intéressant de remarquer que pour les ICM eux-mêmes, le rapport entre la lexie source

[71] « *If someone or* <u>*something*</u> *is* ***responsible*** *for a particular event or situation, they are the cause or they can be blamed for it.* » (Collins Cobuild English Dictionary, 1999 : 1416) (souligné par nous)

et la lexie cible peut se traduire par un changement de classe : ainsi, pour *passage*, la classe de la lexie source est DEPLACEMENT, tandis que pour la lexie cible, elle est plutôt FORME.

Enfin, il est clair que les projections opèrent par l'intermédiaire des classes d'actant, plutôt que par l'intermédiaire des instances actantielles elles-mêmes. C'est ce qui permet, pour un locuteur, de prévoir l'usage d'un ICM avec différentes instances. Nos classes d'actants correspondent ainsi aux cadres conceptuels source et cible (*source domain, target domain*) de Lakoff (1993). Le plus intéressant de cette stratégie d'analyse, aussi imparfaite qu'elle puisse être encore, est qu'elle constitue un pas vers la systématisation de la formation des « noms » de métaphore conceptuelle que Lakoff (1987/2003 ; 1993) énonce sous une forme propositionnelle, du type LES MOLECULES SONT DES PERSONNES.

2.3.2. Conceptualisation des processus biologiques

Deux cas assez complexes sont représentés par les ICM *voie* et *cascade*, qui sont assez fréquents tous les deux (plus de 100 occurrences de *cascade* et plus de 300 occurrences de *voie* dans la totalité du corpus français).

Ces ICM dénotent tous deux une suite d'évènements biologiques, mais sous des modes de conceptualisation légèrement différents :

(16) « Les signaux induits par les facteurs de croissance et les molécules d'adhérence sont transmis au noyau par des relais intracellulaires dont le principal est constitué ***d'une cascade de protéine-kinases*** *nommée «* ***voie de signalisation des MAP-kinases*** *» (…).* » (Charron, 1999 : 1155)

Cascade se retrouve dans des expressions du type :

~ de phosphorylations, ~ d'activations (évènements X)

~ d'enzymes (enzymatique), ~ de protéases, ~ de caspases, ~ de kinases (molécules Y)

~ de signalisation, apoptotique (processus)

La lexie cible *cascade* peut être ainsi définie :

'succession : d'évènements X de même type ou faisant intervenir des molécules X' de même type'

Le processus réalisé (*signalisation, apoptose [apoptotique]*) ne fait pas partie de la définition.

La lexie-source la plus proche a pour sens *'succession : ~ d'évènements X'*. La plus éloignée est celle dont le sens est '*succession de chutes d'eau*'. Par conséquent, les évènements, ici les évènements biologiques, qui se suivent dans le temps sont conceptualisés comme des entités qui se suivent dans l'espace.

Voie apparaît dans des expressions du type :

~ de transmission, ~ de différenciation, ~ de signalisation, ~ de transduction, ~ d'activation, ~ de transformation… (but Y)

~ des MAP-kinases, ~ de l'IP3, ~ des seconds messagers, ~ de l'adényl cyclase…

La lexie cible *voie* peut être ainsi définie :

'succession : d'évènements X ayant un but Y'

La lexie source *voie* la plus proche a pour sens *'suite d'actes X ayant un but Y'* (par exemple, *~ du salut, ~ de perdition, ~ de la connaissance…*), et la plus éloignée a pour sens '*espace allant de X à Y servant à Z'* (*~ de circulation, ~ de service, ~ de communication*).

Dans ce cas, *MAP-kinases, IP3, seconds messagers, adényl cyclase* servent plutôt de nom à la voie. Ce qui est conceptualisé ici, c'est surtout l'ensemble des évènements biologiques qui se

succèdent. La projection entre actants est plus complexe à établir. Ce qui est le plus saillant, c'est une projection qui opère entre *'espace'* et *'succession d'évènements*', et qui revient, comme pour *cascade*, à conceptualiser une suite d'évènements comme un chemin, le temps étant conceptualisé comme un espace. En ce qui concerne le modèle qui peut rendre compte de la CM, c'est la mise en rapport des classes de la lexie source et de la lexie cible qui est la plus évidente, les structures actantielles se prêtant plus ou moins bien à l'exercice.

Ainsi qu'en témoignent les exemples (17) à (19), le réseau lexical est compatible avec la conceptualisation des processus comme des chemins : on relève *aboutir* (qui peut être employé aussi bien dans le cas d'un espace qui se termine [*la route aboutit à la mer*] que pour un fait [*le raisonnement aboutit à la solution*]), *emprunter (les voies)* (*emprunter un chemin, une* route), et *en aval*, qui serait, lui, plus compatible avec l'idée de cours d'eau (ou de cascade), ce qui oriente la conceptualisation, localement, vers les voies fluviales plutôt que les voies terrestres.

(17) « On distingue trois ***voies de signalisation*** *faisant intervenir les MAP-kinases: la voie impliquant les Jun kinases, celle de la p38 kinase, et celle qui* ***aboutit*** *à la phosphorylation des kinases ERK (*extracellular regulated kinase*) par une MAP kinase kinase, nommée MEK (*mitogen extracellular signal kinase*), située* ***en aval*** *de Raf et de Ras. »* (Charron, 1999 : 1155)

(18) « La transmission du signal engendré par diverses cytokines, lorsqu'elles se fixent sur leur récepteur, ***emprunte*** *également ce type de* ***voie*** *directe. »* (Étienne, 1999 : 187)

(19) « Dans ce cas le médicament ***emprunte*** *les* ***voies*** *de transformation chimique. »* (Bourin, 1994 : 56)

Du point de vue de l'analyse des ICM, on remarquera qu'il est plus facile de percevoir le mode de conceptualisation à partir de *aboutir*, *emprunter* et *en aval*, car on retombe dans une situation où il est plus aisé de mettre en correspondance des dasses d'actants prototypiques (les actants prototypiques de *en aval* dénotent des cours d'eau, ceux de *aboutir* des chemins ou des actions). On peut aussi, à partir de ces exemples, appréhender la complexité des interrelations sémantiques qui finissent par mener à une sorte de jeu de miroirs à l'infini.

2.3.3. Représentations fictives semblables et conceptualisation réciproque

Considérons les exemples suivants :

(21) « L'artère circonflexe humérale postérieure (…) ***irrigue*** *le deltoï de, le chef long du triceps et le chef latéral. » (Chevallier 1998 : phr 38)*

(22) « [la veine basilique] (…) ***se jette*** *soit dans les veines brachiales, soit dans la veine axillaire. » (Chevallier, 1998 : phr 80)*

(23) «Le ***confluent*** *veineux suboccipital* ***donne naissance*** *à la veine vertébrale et à la veine jugulaire postérieure. » (Chevallier 11998 : phr 103)*

(24) «Ce ***confluent*** *est* ***drainé*** *par trois voies : la veine jugulaire externe; la communicante intraparotidienne, qui, après un* ***trajet*** *intraglandulaire, sort de la parotide près du digastrique,* ***traverse*** *la cloison interparotidomaxillaire,* ***longe*** *le pôle postérieur de la sous-maxillaire et* ***va se jeter*** *dans la veine faciale; la veine carotide externe, toujours irrégulière et peu nette, qui* ***suit*** *exactement* ***le trajet*** *de l'artère carotide externe et* ***se jette*** *dans la jugulaire interne au voisinage du tronc thyro-linguo-facial. » (Grégoire 1991 : phr 718)*

Le réseau lexical mis en évidence dans les exemples (21) à (24) s'applique à la fois aux vaisseaux sanguins et aux cours d'eau, ce qui induit, dans les deux cas, une représentation mentale de déplacement fictif. Ce type de conceptualisation s'applique de façon générale aux

chemins, terrestres ou fluviaux. Inversement, les grandes rues des villes sont appelées «artères », auxquelles s'applique également le déplacement fictif :

*(25) « Principale **artère** du Plateau Mont-Royal, l'avenue du Mont-Royal **traverse**, de l'ouest vers l'est, ce célèbre quartier appelé simplement «le Plateau». »* (Séguin 2001, sp)

Ce qui permet de proposer le modèle suivant (Figure 1) :

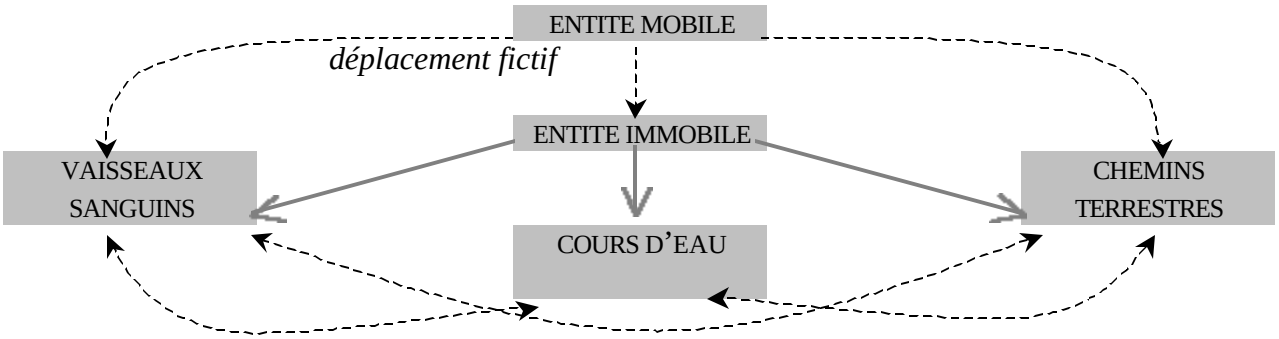

Figure 1 – Projections réciproques entre classes d'entités représentées de façon analogue

Les conceptualisations ne sont cependant pas parfaitement bijectives, car elles sont contraintes par la cible (« *Target domain overrides* » (Lakoff, 1993 : 216)) : par exemple,**la rue irrigue la ville* est invalide, tandis que *l'artère irrigue le muscle* et *le fleuve irrigue la plaine* sont parfaitement admissibles. La restriction provient de la fonction de l'artère et du fleuve, qui est d'amener le sang ou l'eau dans des territoires anatomiques ou géographiques, réciproquement. Un autre de type de restriction peut venir du deuxième actant, comme dans le cas de *se jeter* : ainsi **le fleuve se jette sur la place, *la veine se jette dans le muscle* sont invalides[72], mais *le fleuve se jette dans le lac, la veine X se jette dans la veine Y,* sont admissibles : la condition imposée par la structure actantielle de *se jeter* est que le deuxième actant dénote un espace contenant un liquide, de l'eau (*mer, fleuve*) ou du sang (*veine*). Cette condition est sans doute héritée de la lexie source de *se jeter* exprimant le saut d'une personne X dans quelque chose Y contenant un liquide (*piscine, cours d'eau, lac, mer*…). C'est cette restriction qui valide l'interférence cognitive entre les vaisseaux sanguins et les cours d'eau. Par contre, lorsqu'un ICM s'applique aussi bien aux vaisseaux, aux chemins qu'aux cours d'eau sans restriction, aucune interférence particulière n'est saillante, hormis le déplacement fictif qui s'applique aux trois situations : *la rue traverse la ville, l'artère traverse le muscle, la rivière traverse la plaine*.

3. Saillance des conceptualisations métaphoriques et réseaux lexicaux

Un mode de conceptualisation ne devient conventionnel que s'il est appuyé par un réseau lexical suffisamment riche partagé par les locuteurs. Identifier le moment à partir duquel ceci se produit relève de la psychologie cognitive, mais le nombre des ICM, ainsi que leur fréquence et leur répartition dans différents textes sont autant de paramètres permettant d'évaluer l'originalité ou la banalité d'un mode de conceptualisation particulier. Le renforcement mutuel des ICM dans un texte ou un domaine fait émerger, au plan cognitif, le mode de conceptualisation. On constate ainsi dans le tableau 4 que les actants typiques des lexies sources relèvent le plus souvent de la classe des êtres vivants et dans certains cas, de celle des êtres humains, et que les lexies cibles,

[72] La formulation correcte est *la veine draine le muscle*.

elles, relèvent essentiellement de la classe des molécules ou des cellules. Le même phénomène est observé dans le cas des ICM multiactantiels. Si un ICM est isolé, il sera perçu comme un hapax, une expression métaphorique « déviante », au mieux une figure style, au pire une incongruité. Deux paramètres sont envisagés : la diversité et la cohérence lexicales.

3.1. Diversité des ICM et organisation hiérarchique des classes d'actants des lexies sources

Un des paramètres du renforcement d'une CM résulte, au plan cognitif, de la diversité d'ICM cohérents. Le degré de saillance d'un mode de conceptualisation donné est en rapport avec le réseau lexical exprimé. Ainsi, un certain nombre d'ICM évoquent une conceptualisation des molécules comme des êtres humains :

communauté, coloniser, mort, suicide, parenté, partenaire, famille, population…
migrer, coopérer, agir, intervenir…

Toutefois, certains ICM peuvent aussi s'appliquer à la classe des ANIMAUX$_{(2)}$[73] : *coopérer, migrer, population*… Par conséquent, la projection métaphorique opère aussi depuis le niveau ANIMAL$_{(1)}$. Enfin, la conceptualisation peut procéder d'un niveau plus élevé, celui des êtres vivants : *mort, coloniser, jeune, immature…*

La conceptualisation peut ainsi devenir relativement floue, lorsque les classes d'actants prototypiques des lexies sources relèvent de catégories organisées hiérarchiquement (Figure 2).

Selon Lakoff (1993 : 211), les projections métaphoriques opèrent à partir des catégories superordonnées : il se pourrait que les projections se produisent plutôt à partir de différents niveaux pour aboutir à une intégration conceptuelle, laquelle devient plus saillante à un niveau donné. La façon dont le niveau le plus saillant se détermine reste à déterminer : ce pourrait être soit le plus bas, soit celui auquel s'appliquent le plus d'ICM.

Les projections métaphoriques sont partielles : on peut dire *jeune cellule*, *cellule immature*, *molécule immature*, mais pas *jeune molécule*. Les notions auxquelles renvoient *jeune cellule* et *cellule immature* sont d'ailleurs différentes, la première concernant l'âge de la cellule, le deuxième, son état fonctionnel. Comme les deux notions sont corrélées (les cellules immatures sont généralement jeunes), elles pourraient être confondues à tort. Par ailleurs, le fait que *jeune molécule* ne se dise pas n'implique pas nécessairement une différence de conceptualisation : peut-être que la question de «l'âge » d'une molécule n'est tout simplement pas pertinente. La réponse à cette question n'est plus d'ordre linguistique ou cognitif, mais de nature scientifique.

Enfin, la projection n'opère pas nécessairement des classes les plus élevées dans la hiérarchie vers les plus basses. Une famille est ensemble de personnes apparentées biologiquement, tandis qu'une famille de molécules est un ensemble de molécules « apparentées » par leur structure chimique, et une famille d'êtres vivants (dans les taxonomies) est un ensemble d'êtres vivants « apparentés » en raison de certaines caractéristiques biologiques partagées (morphologiques ou, dans les nouvelles nomenclatures, génétiques). Par conséquent, il est probable que la projection opère depuis la classe HUMAIN sur la classe ETRE VIVANT, et non l'inverse.

[73] ANIMAL$_{(1)}$ s'oppose à VEGETAL, tandis que ANIMAL$_{(2)}$ s'oppose à ETRE HUMAIN.

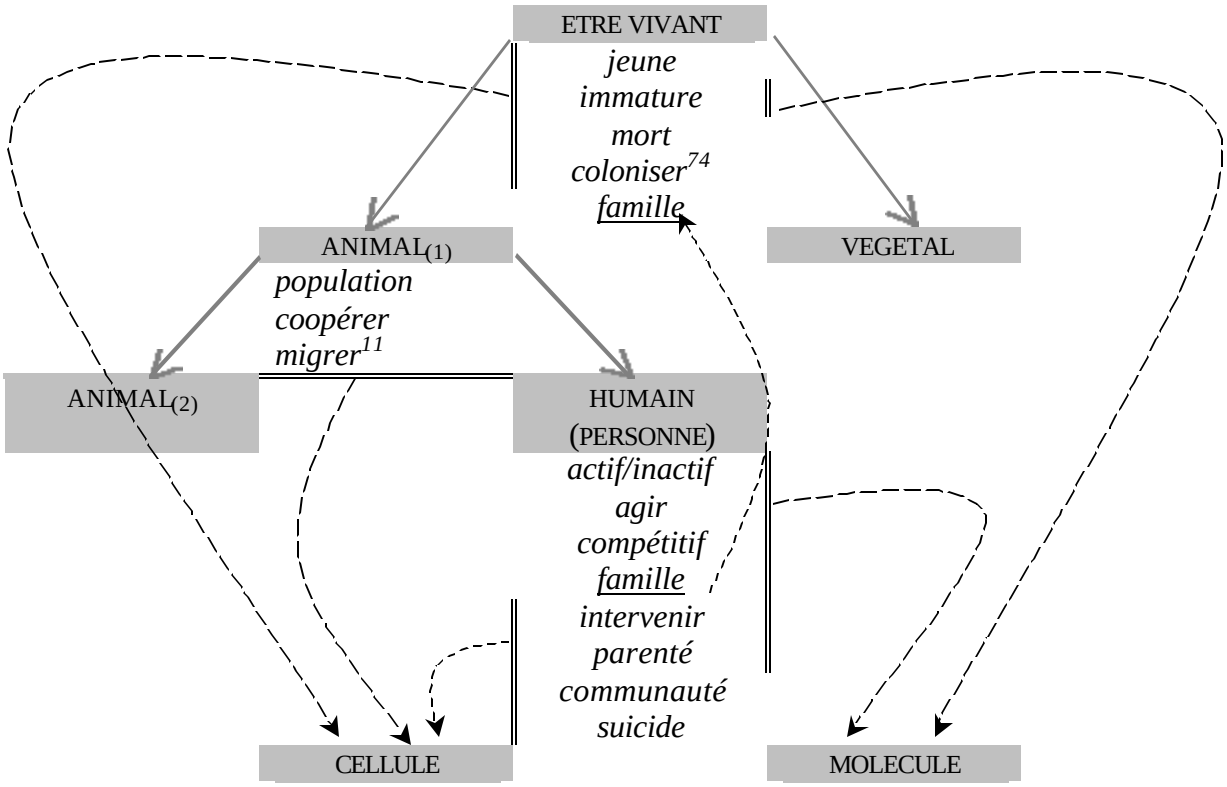

Figure 2 – Conceptualisation métaphorique des cellules et des molécules

3.2. Cohérence lexicale

La cohérence lexicale peut être décrite sur deux axes : paradigmatique et syntagmatique.

3.2.1. Cohérence lexicale paradigmatique

Dans l'exemple (26), les ICM sont cohérents avec la conceptualisation des cellules comme des personnes ou des animaux$_{(2)}$ se déplaçant dans une région géographique : *région, territoire, migrer, coloniser.*

(26) « Les cellules des crêtes neurales céphaliques sont issues de l'épithélium neural, et ***migrent*** *vers la* ***région*** *ventrale de l'embryon où elles* ***colonisent*** *différents* ***territoires*** *pour former, entre autres, les structures de la face (muscles et cartilages). »* (Alfandari 1999, 1151)

Or, *région* et *territoire* représentent des noms d'actants typiques des lexies sources *migrer* et *coloniser*. Par conséquent, le lien sémantique existant entre ces ICM renforce deux modes de conceptualisation complémentaires et cohérents, l'un, des cellules conceptualisées comme des personnes ou des animaux, l'autre, l'organisme comme un espace géographique.

3.2.2. Cohérence lexicale syntagmatique : collocations

Un phénomène particulièrement intéressant, notamment au plan de l'idiomaticité, concerne le « transfert » de collocatifs. Ainsi, les collocations *emprunter un chemin, une route, une voie de circulation* se trouvent transposées pour l'ICM *voie* en biologie cellulaire :

[74] Voir plus bas exemple (26).

(27) « Les séquences de tri, d'adressage et de rétention des protéines ***empruntant la voie de biosynthèse/sécrétion et d'endocytose*** *interagissent avec les protéines de manteau des vésicules de transport (vésicules recouvertes de clathrine, vésicules COP...) »* (Goud, 14, 1338)

(28) « Une idée très séduisante pour améliorer cette étape consiste à utiliser les propriétés des peptides NLS afin ***d'emprunter les voies cellulaires physiologiques*** *du transport nucléaire. »* (Behr et Belguise-Valladier, 1999 : 757)

Le même phénomène est observé avec le collocatif *membre*, qui est accompagné par la base *famille* dans quasiment tous les cas sur une centaine de collocations relevées :

(29) « La ß-arrestine-1 est un ***des membres de la famille des arrestines****, découvertes pour leur capacité d'interagir avec les RCPG sous leur forme phosphorylée par des protéine-kinases spécifiques, les kinases des RCPG ou GRK. »* (Bouvier et Angers, 1999 : 741).

3.3. Intégration conceptuelle

Lorsqu'un ICM est exprimé en contexte, la simultanéité des représentations fictives et factives induites amène la perception de « dissonance cognitive », qui se résoud cependant en une intégration conceptuelle (Fauconnier et Turner, 1998) menant à la compréhension de l'énoncé. Lorsque l'intégration conceptuelle ne se fait pas, ou se fait mal, le résultat peut être, selon le cas, un énoncé étrange, absurde (pouvant être à la source d'effets humoristiques) ou incompréhensible (pouvant être à l'origine d'erreurs de sens en traduction).

Dans quelques cas, assez rares, la cohérence syntagmatique n'est pas respectée :

(30) « Les colonocytes expriment des récepteurs apicaux et basolatéraux pour le transport des acides aminés, mais les ***mécanismes*** *de la signalisation intracellulaire* ***empruntés*** *par la glutamine n'ont pas encore été définis. »* (Ruemmele, 1999 : 52)

(31) « Les ***membres de la famille*** *de Bcl-2 et une cascade de protéases à activité cystéine nommées caspases sont des* ***effecteurs principaux de la machinerie apoptotique****, présente dans toutes les cellules. (…) Akt phosphoryle de façon directe* ***deux membres de la machinerie apoptotique*** *: la caspase 9 et BAD, une protéine de la famille de Bcl-2. »* (Brunet 1999 : 897)

Un cas intéressant d'incohérence paradigmatique a été relevé, dans lequel la structure actantielle de *emprunter* a été inversée :

(32) « La deuxième ***voie de transmission du signal*** *n'****emprunte*** *pas de* ***molécules « messagers »*** *mais implique des cascades de phosphorylation mises en route par l'activation de récepteurs qui ne traversent qu'une fois la membrane. »* (Pecker, 1998 : 1010)

Toutefois, la règle générale est que plusieurs modes de coexistent harmonieusement (au point où l'on ne s'en rend plus nécessairement compte, si le domaine est familier !) :

(33) « Les ***membres de la famille*** *de Bcl-2 et une* ***cascade*** *de protéases à* ***activité*** *cystéine nommées caspases sont des effecteurs principaux de la* ***machinerie*** *apoptotique, présente dans toutes les cellules. »* (Brunet 1999 : 897)

Nous faisons l'hypothèse que cette intégration conceptuelle se construit et se remodèle au fur et à mesure de l'acquisition des connaissances, que ce soit au cours de l'apprentissage ou de la survenue des découvertes, et que cela a des conséquences non seulement sur des apprentissages complexes tels que celui des langues et de la traduction, mais aussi sur l'activité scientifique elle-même (nous pensons notamment à la difficulté avec laquelle les chercheurs eux-mêmes ont admis le concept de « suicide cellulaire », tant l'idée de vie est liée au développement et à la multiplication des cellules (Almeisein, 2003)).

Conclusion

Le présent travail fait état d'une méthode d'analyse et d'un ensemble de résultats qui permettent de cerner certains éléments clés de la conceptualisation métaphorique en sciences. Outre la coexistence de plusieurs modes de conceptualisation, différents éléments ont été identifiés : le rôle des actants et des classes d'actants, l'interaction réciproque entre différents cadres conceptuels, l'importance des réseaux lexicaux et de la cohérence lexicale paradigmatique et syntagmatique, l'organisation hiérarchique des classes d'actants des lexies sources et cible, l'intégration conceptuelle. Le caractère indispensable de la conceptualisation métaphorique (même si on peut le regretter (Gaudin, 1998)) se traduit par le fait que dans nombre de cas, avoir recours à une expression induisant d'emblée une représentation factive est pratiquement impossible. La conceptualisation métaphorique est en fait un procédé économique dont l'intelligence s'accommode parfaitement. Par ailleurs, les données recueillies plaident contre un découpage strict entre une langue de spécialité et la langue commune et/ou d'autres langues de spécialité. En effet, dans nombre de cas, les ICM sont indispensables pour l'idiomaticité en raison de leur implication conceptuelle, mais ils n'ont pas de sens spécialisé exclusif (par exemple les verbes induisant une représentation de déplacement fictif).

Il faut maintenant approfondir l'étude des différences entre l'anglais et le français, ce qui se révèlera crucial pour les applications en traduction. Dans cette perspective, il sera intéressant de revisiter la question des interférences linguistiques et de l'équivalence, laquelle devrait prendre en compte les représentations conceptuelles dans les langues en présence, avec les réseaux lexicaux correspondants. Il se pourrait qu'une des différences majeures entre traducteurs débutants et expérimentés soit l'acquisition (plus ou moins conscientisée) des modes de conceptualisation, se traduisant par une idiomaticité accrue. Nous pensons de plus que ce type d'approche a de nombreuses applications : représentation des connaissances, dictionnairique, rédaction, apprentissage des langues et acquisition des connaissances spécialisées.

Dans les recherches futures, il sera également important d'aborder la conceptualisation métonymique. On peut de plus se poser la question de savoir comment concilier les représentations terminologiques, lexicales et cognitives dans des ouvrages *ad hoc*, et d'ailleurs les méthodes d'annotation permettraient de générer des « dictionnaires dynamiques » facilitant le repérage de solutions de traduction. Enfin, la méthode d'annotation appliquée en diachronie pourrait permettre d'étudier l'évolution des représentations cognitives dans un domaine particulier.

Remerciements

Nous remercions le Conseil de recherche en sciences humaines du Canada et le Fonds québécois de la recherche sur la société et la culture pour leur soutien financier.

Bibliographie

ALMEISEN J.-C., 2003, *La sculpture du vivant – Le suicide cellulaire ou la mort créatrice*, Coll. Points-Sciences, Seuil, Paris.

BOUVERET M., 1998, « Un cas de métaphore : créativité linguistique et rôle en innovation biotechnologique », dans *La mémoire des mots, Actes des V^e Journées Scientifiques du*

Réseau Thématique LTT, Tunis, 25-27 septembre 1997, AUPELF-UREF/Serviced, pp. 315-326.

CHARTERIS-BLACK J., 2004, *Corpus Approaches to Critical Metaphor Analysis*, Palgrave MacMillan, New York.

COLLOMBAT I., 2003, « Le discours imagé en vulgarisation scientifique : étude comparée du français et de l'anglais », *metaphorik.de 05/2003*, http://www.metaphorik.de/francais.htm

COLLINS COBUILD ENGLISH DICTIONARY, 1999, dir. par J. Sinclair, HarperCollins Publishers, Londres.

DE VILLERS M.-É., 1997, *Multidictionnaire de la langue française*, Québec-Amérique, Montréal.

ECO U., 1988/2006, *Sémiotique et philosophie du langage*, Presses Universitaires de France, Paris.

DEIGNAN A., 2005, *Metaphor and Corpus Linguistics*, John Benjamins, Amsterdam/Philadelphia.

DURY P., 1999, « Variations sémantiques en terminologie : étude diachronique et comparative appliquée à l'écologie », dans *Sémantique des termes spécialisés*, V. Delavigne et M. Bouveret, Coll. Dyalang, Université de Rouen, CNRS, n°273, pp. 17-33.

DUVIGNAU K., 2002, *La métaphore, berceau et enfant de la langue*, Thèse présentée devant l'Université de Toulouse II.

FAUCONNIER G. et TURNER M. 1998, *The way we think : conceptual blending and the mind's hidden complexities*, Basic Books, New York.

FOX KELLER E., 1999, *Le rôle des métaphores dans les progrès de la biologie*, Coll. Les empêcheurs de tourner en rond, Institut Sanofi-Synthélabo, Paris.

GAUDIN F., 1998, « Métaphores et diachronie dans les sciences : le cas de code, patrimoine, sélection », dans *La mémoire des mots, Actes des V^e Journées Scientifiques du Réseau Thématique LTT*, Tunis, 25-27 septembre 1997, AUPELF-UREF/Serviced, pp. 243-250.

KOLLER V., 2004, *Metaphor and Gender in Business Media Discourse – A Critical Cognitive Study*, Palgrave MacMillan, New York.

KORDON C., 1991, *Le langage des cellules*, Coll. Questions de science, Hachette, Paris.

LAKOFF G. et JOHNSON M. 1980/2003, *Metaphors We Live By – With a New Afterwords*, The University of Chicago Press, Chicago.

LAKOFF G., 1993, « The contemporary theory of metaphor », in : *Metaphor and thought*, 2^e édition, dir. par A. Ortony, Cambridge University Press, Cambridge, p. 203-251.

LAKOFF G. et TURNER M., 1989, *More than Cool Reason: A Field Guide to Poetic Metaphor*, University of Chicago Press, Chicago.

LUBIN L., 2006, *Étude des métaphores conceptuelles utilisées dans la description des structures anatomiques*. Mémoire de maîtrise de l'Université de Montréal, Département de linguistique et de traduction.

LE NY J.-F., 1979, *La sémantique psychologique*, Coll. Le psychologue, Presses Universitaires de France, Paris.

LE NY J.-F., 2005, *Comment l'esprit produit du sens*, Éditions Odile Jacob, Paris.

NEWMARK P., 1981, « The translation of metaphor », *The Incorporated Linguist – The Journal of the Institute of Linguists*, vol. 20, n°1, pp. 49-54.

MEL'ÈUK I.A., CLAS A. et POLGUÈRE A., 1995, *Introduction à la lexicologie explicative et combinatoire*, Duculot / Aupelf - UREF, Louvain-la-Neuve.

MIROWSKI P., 2001, *Plus de Chaleur que de Lumière*, Economica, Paris.

OLIVEIRA I., 2006, « La métaphore terminologique sous un angle cognitif », Meta, vol. 50, n°4, dans *Actes du 50e anniversaire de Meta – La traduction proactive,* publié sous forme de CD-rom.

PERLERIN V., FERRARI S., et BEUST P., 2002, « Métaphores et dynamique sémique : expériences sur corpus », Actes des 2èmes Journées de la Linguistique de Corpus, Lorient (http://users.info.unicaen.fr/~ferrari/recherche/publis.html).

SCHAEFFNER C., 2004, « Metaphor ans translation : some implications of a cognitive approach », *Journal of Pragmatics*, vol. 36, p. 1253-1269.

STAMBUK A., 1998, « Metaphor in Scientific Communication », Meta, vol. 43, n°3, pp. 373-379.

TABAKOWSKA E., 1993, *Cognitive Linguistics and Poetics of Translation*, Language in Performance, coll. dir. par W. Hüllen et R. Schülze, Gunter Narr Verlag, Tübingen.

TAMBA I., 1981, *Le sens figuré*, Presses Universitaires de France, Paris

TAMINE J., 1978, *Description syntaxique du sens figuré : la métaphore*, Thèse de doctorat d'état, Paris VII.

TEMMERMAN R., 2000, *Towards New ways of terminology Description : The sociocognitive Approach*, John Benjamins, Amsterdam.

TEMMERMAN R., 2002, « Metaphorical models and the translation of scientific texts », *Linguistica Antverpiensa*, vol. 1, pp. 211-226.

TEMMERMAN R., 2006, « Sociocultural situatedness of terminology in the life sciences: The history of splicing », à paraître dans *Body, language and mind. Vol. II. Interrelations between Biology, Linguistics and Culture*, F. Roslyn, J. Zlatev et T. Zieke, Springer Verlag, Tübingen.

TESNIÈRE L., 1965, *Éléments de syntaxe structurale*, 2e édition revue et corrigée, Klincksieck, Paris.

TALMY L., 2001, « Toward a cognitive semantics », *Volume I: Concept structuring systems*, The MIT Press, Cambridge.

VANDAELE S., 2000, « Métaphores conceptuelles et traduction biomédicale. », *La traduction : théorie et pratiques, actes du colloque international Traduction humaine, traduction automatique, interprétation*, sous la direction de S. Méjri, T. Baccouche, A. Clas. G. Gross, Tunis, 28-29 septembre 2000, Publications de l'ENS, p. 393-404.

VANDAELE S, 2003, « Métaphores conceptuelles et traduction médicale », *TTR*, (XV) 1, p. 223- 239.

VANDAELE S., 2004, « Deciphering metaphorical conceptualization in biomedicine: towards a systematic analysis » In : *New Directions in LSP studies, Proceedings of the 14th European Symposium on Language for Special Purposes*, 18 - 22 août 2003, dir. par M. Rogers et K. Ahmad, p. 195-202.

VANDAELE S., 2005, « Métaphores conceptuelles et fonctions lexicales : des outils pour la traduction médicale et scientifique », *Actes du IIIe congrès international de traduction spécialisée*, Barcelone, Université Pompeu Fabra, 4 - 6 mars 2004, p. 275-286.

VANDAELE S et LUBIN L., 2005, « Approche cognitive de la traduction dans les langues de spécialité : vers une systématisation de la description de la conceptualisation métaphorique », *META*, numéro spécial dirigé par H. Lee-Jahnke, vol. 20(2), p. 415-431.

VANDAELE S. et BOUDREAU S., 2006, « Annotation XML et interrogation de corpus pour l'étude de la conceptualisation métaphorique », *Actes des 8e Journées internationales d'analyse statistique des données textuelles (JADT2006)*, Université de Besançon, 19-21 avril 2006, p. 951-959.

VIDAL V. et CABRÉ M. T., 2006, « La combinatoria léxica especializada : combinaciones metafóricas en el discurso de Genoma Humano », à paraître.
van RIJN-van TONGEREN G. W., 1997, *Metaphors in Medical Texts,* Editions Rodopi, Amsterdam/Atlanta.
YU N., 1998, *The Contemporary Theory of Metaphor : A Perspective from Chinese,* John Benjamins, Amsterdam/Philadelphia.

Bibliographie des textes cités en exemple

Les références précédées d'un astérisque renvoient à des textes ne faisant pas partie des corpus annotés.

ALBERTS B, BRAY D. et coll., 1998, *Essential Cell Biology - An Introduction to the Molecular Biology of the Cell*, Garland Publishing, New York.
ALFANDARI D., COUSIN H. et coll., 1999, « Les protéines de la famille ADAM : protéolyse, adhérence et signalisation », *Médecine sciences*, vol. 15(10), p. 1148-1151.
ALLIET J. et LALÉGÉRIE P., 1997, *Cytobiologie*, Ellipses, Paris.
BAR I. et GOFFINET A.-M., 1999, «Récepteurs des lipoprotéines et signalisation par le Reelin au cours du développement cérébral », *Médecine sciences*, vol. 15(11), p. 1284-1285.
BEHR J. P. et BELGUISE-VALLADIER P., 1999, «Les signaux de localisation nucléaire : un sésame cellulaire pour le transport d'ADN ? », *Médecine sciences*, vol. 15(5), p 757-758.
BOURIN, M., 1994, *Pharmacologie générale et pratique*, 2e éd., Ellipses, Paris.
BOUVIER M. et ANGERS S., 1999, « Nouveaux échafaudages protéiques modulaires pour les récepteurs couplés aux protéines G », *Médecine sciences*, vol. 15(5), p. 741-743.
CHARRON J. et L. JEANNOTTE, 1999, « Le rôle essentiel de MEK1 lors de l'angiogenèse placentaire », *Médecine sciences*, vol. 15(10), p. 1155-1157.
CHEVALLIER J.-M., 1998, Anatomie, Appareil locomoteur, Médecine-Sciences Flammarion, Paris.
ÉTIENNE J. 1999, *Biochimie génétique - Biologie moléculaire*, Masson, Paris.
*FARZAN M., CHOE H., MARTIN K., et coll. (1997) « Two Orphan Seven-Transmembrane Segment Receptors Which Are Expressed in CD4-positive Cells Support Simian Immunodeficiency Virus Infection » *Journal of Experimental Medicine* (186), vol. 3, p. 405-411.
GOUD B., 1999, «Le code d'adressage des protéines », *Médecine Sciences*, vol. 15(11), p. 1336-1338.
GRAY, H. 1989 *Gray's Anatomy*, dir. par P.L. Williams et coll., 37e édition, Livingstone, New York.
GRÉGOIRE C. et coll., 2004, *Précis d'anatomie*, 11e édition, Technologie et documentation, Paris.
LESCA P. et T. PINEAU, 1999, « Toxicité de la dioxine : rôle des protéines PAS (Étapes de transmission du signal) », *Médecine Sciences*, vol. 15(12), p. 1379-1387.
PECKER F. et coll., 1998, «Le rôle messager de l'acide arachidonique dans le cardiomyocyte », *Médecine sciences*, vol. 14(10), p. 1009-1016.
*ROCKMAN H. A., KOCH W. et J. LEFKOWITZ R.J., 2002, « Seven-transmembrane-spanning receptors and heart function », *Nature*, vol. 415, p. 206-212.

ROUVIÈRE H. et DELMAS A., 1991, *Anatomie humaine : descriptive, topographique et fonctionnelle*, 13e éd. revue et augmentée, Masson, Paris.

RUEMMELE F. et coll., 1999, « Les mécanismes moléculaires de la régulation du renouvellement des cellules épithéliales intestinales par les nutriments », *Gastroentérologie Clinique et Biologique*, vol. 23., p. 47-55.

*SÉGUIN Y., 2001, *Randonnée pédestre – Montréal et environs*, Guide de voyage Ulysse, http://www2.canoe.com/voyages/sechapper/archives/2005/03/20050324-112609.html

COMPTE RENDU

Marie-Madeleine Bertucci, Violaine Houdart-Merot (dirs.), 2005 :
Situations de banlieues, Enseignement, langues, cultures, Edition de l'Institut National de Recherche Pédagogique, collection Education, Politiques, Sociétés, Lyon, 290 pages, ISBN 2-7342-1013-4.

Véronique Miguel

Université de Rouen – FRE 2787 DYALANG

Cet ouvrage de 290 pages intitulé *Situations de banlieues, Enseignement, langues, cultures* et publié par l'INRP, se compose de 26 articles et contributions rédigés par 32 auteurs, sous la direction de Marie-Madeleine Bertucci et Violaine Houdart-Merot. Les parcours de ces deux chercheurs montrent qu'elles sont particulièrement sensibles aux questions du langage et de la littérature, en termes de variations et d'intertextualité. L'université à laquelle elles sont rattachées (Cergy-Pontoise) fait des banlieues leur terrain de recherche privilégié. Mais le langage est avant tout envisagé comme une partie d'un tout social complexe, qui s'exprime de façon hétérogène dans un milieu marqué d'abord par la diversité. L'objectif de ce travail est de poser une réflexion plurielle sur les situations de banlieues, postulant que c'est la diversité des regards qui assure la qualité de l'observation et de la réflexion. Cette approche méthodologique se présente comme garante des pistes et propositions didactiques qui sont faites à la faveur de la dernière partie de l'ouvrage.

Ces articles, qui trouvent leur origine dans un colloque ayant eu lieu en novembre 2004, à l'université de Cergy-Pontoise, sont le fruit d'une réflexion construite par un ensemble de chercheurs en linguistique ou en géographie, de membres de l'Education nationale, de psychanalystes, d'acteurs de la vie culturelle…. Et en contrepoint des analyses souvent complétées par une bibliographie proposée par les auteurs, se glissent des comptes-rendus d'expériences qui offrent un « focus » sur une situation particulière, permettant d'éclairer de façon pragmatique et didactique la démarche théorique. Une tentative de définition des banlieues est faite (*Regards croisés sur les situations de banlieues*) après une introduction qui pose le cadre de la réflexion. Un regard particulier est posé sur les *Pratiques culturelles et créations littéraires* dans une deuxième partie, pour aboutir à une réflexion sur *Enseigner en situation de banlieue*. La lecture de cet ouvrage aux voix plurielles est donc utile tant aux différents intervenants de l'enseignement en situation de banlieue qu'aux chercheurs ou aux acteurs sociaux. Il s'adresse de façon générale à des lecteurs qui désirent ne pas s'arrêter à des simplifications médiatiques sur une situation constitutive de notre société actuelle.

Dans une introduction conséquente qui cherche à mettre en perspective banlieue et école, Marie-Madeleine Bertucci et Violaine Houdart-Merot soulignent le fait qu'on ne peut réfléchir à la question des banlieues qu'en se distançant des stéréotypes péjoratifs et abusivement simplistes véhiculés par les médias. L'étude de textes littéraires et des répertoires langagiers est un moyen privilégié pour dégager les contenus et la force de ces représentations. On y retrouve les thèmes récurrents de l'exclusion, de la violence… dans le cadre de trois caractéristiques majeures des regards portés sur les « milieux difficiles » : précarité et marginalisation, regard ethnicisant, enlisement dans le manque d'autonomisation. La violence potentielle est alors omniprésente. Ce prisme de lecture est handicapant pour les habitants et ne leur permet pas de trouver une place dans la « cité » en tant que citoyens à part entière, selon le modèle français républicain. Voilà pourquoi le terrain de l'école est d'un enjeu capital. Au lieu de permettre l'intégration, la norme scolaire semble parfois stigmatiser les différences car elle ne légitime pas la diversité reconnue dans le milieu des banlieues. Ce malentendu structurel repose sur un paradoxe de l'école qui, selon les auteurs, cherche par nature à intégrer en gommant les différences. Des pistes sont données pour dépasser ces disfonctionnements : rencontre et prise en compte de la diversité culturelle et langagière, dans le respect de l'altérité.

Pour les auteurs, vouloir rendre compte des banlieues, que les médias s'accordent le plus souvent à stigmatiser, c'est donc d'abord chercher à leur reconnaître une réalité multiple, aux facettes tant géographiques que culturelles, éducatives, historiques, psychologiques… Dans une première partie de 84 pages, 10 articles aident à définir les banlieues, tout en mettant en évidence leur caractère hétérogène. C'est cette démarche que les auteurs nomment *Regards croisés sur les situations de banlieues*. Il faut comprendre ici un effort réel de regards pluriels, et une attention particulière aux réalités multiples que le concept flou de « banlieue » véhicule (voir l'article de Hervé Vieillard-Baron : « La banlieue au risque des définitions » et celui de Pierre Zembri : « Les nouvelles périphéries urbaines : pour une relativisation de la notion classique de banlieue ») : il s'agit à priori tout autant de villes nouvelles que de zones pavillonnaires ou de cités à loyers modérés. Si les quartiers périphériques se comprenaient encore au dix-neuvième siècle comme dépendants d'un réel centre urbain, les zones périurbaines s'organisent aujourd'hui autour de pôles créés artificiellement pour faciliter les transports, les échanges économiques et les services. Une vie indépendante des villes se développe donc, avec toutes ses composantes ; on peut parler de « mosaïques urbaines ». Il est trop facile – et faux – de penser en terme d'ethnicisation des quartiers qui se définissent essentiellement par leur diversité, tant linguistique que culturelle. Françoise Lorcerie le souligne dans son article « Culture, ethnicité, identité. Repenser l'approche interculturelle », ainsi que Muriel Molinié dans « Regards sur le plurilinguisme en banlieue ». Josiane Frossart complète cette approche avec son regard de psychanalyste dans « Exil et bilinguisme. Quand la langue maternelle est en question ». Une grille de lecture éloignée des réalités mène donc souvent à des représentations (voire autoreprésentations) fondées sur un fort sentiment d'injustice chez les jeunes (Valérie Caillet), une crainte réelle de l'insécurité et de la violence (Didier Desponds)… Certaines actions sociales et éducatives relatées ici montrent cependant qu'on peut éviter ces stigmatisations en tenant compte des particularités. On lira à ce sujet l'article de Véronique Bordes sur « les Fêtes de banlieue », mais aussi les expériences positives menées par certaines instances de l'Education nationale (« Les adolescents décrocheurs », de Corinne Tyzler, «La prise en compte de la violence en milieu scolaire dans l'académie de Versailles », de Bruno Robbes).

Cette approche à la fois généralisante et plurielle trouve un écho un peu plus précis dans la deuxième partie (6 articles, 73 pages), qui se spécialise dans un regard porté sur la culture et la littérature : *Pratiques culturelles et créations littéraires*. On pourra ici encore apprécier le choix du pluriel. Elisabeth Auclair montre dans « Offres et demandes culturelles, ou la spécificité du développement culturel en banlieue » que tout en reconnaissant l'émergence de cultures diverses dans les quartiers périphériques, la tentation est grande de les différencier de ce que certains sociologues identifient comme la culture « légitime », ou culture « classique », accessible essentiellement dans les centres urbains. Même si les écrivains issus des banlieues refusent le simplisme des classifications, cette différenciation se retrouve souvent dans la littérature, comme le montre Christiane Chaulet-Achour dans « Banlieue et Littérature ». Certains ouvrages pour la jeunesse continuent à véhiculer ces représentations fausses, d'autres présentent des milieux plus proches des réalités (voir l'étude faite par Max Bulten dans « Images des banlieues dans les ouvrages recommandés à l'école et au collège »). Mais s'arrêter à ce constat serait méconnaître le génie des textes de Marguerite Duras qui fait de la banlieue moins un objet de création qu'un lieu créateur par essence (voir l'article de Simona Crippa). L'article de Serge Martin « Au-delà des banlieues il y a des hommes libres » trouve alors toute sa pertinence : c'est la rencontre qui rend libre, la rencontre physique, mais aussi littéraire. C'est là que se tisse l'esprit citoyen. Se profile alors très concrètement l'expérience des Padox, marionnettes humaines qui vont à la rencontre de ceux qui vivent dans la cité (voir l'article de Dominique Houdart), et que le lecteur peut aussi deviner sur la première de couverture.

Le lieu quotidien de la rencontre reste l'école, et l'on comprend qu'une réflexion sur ***Enseigner en situation de banlieue*** ne peut être constructive qu'à la lumière de ce qui a été analysé dans les deux premières parties. Cette troisième et dernière partie regroupe 10 articles dans 93 pages, elle n'étudie que l'enseignement du français, étant entendu que dans la classe de français se déclinent difficultés et remédiations tout à la fois : c'est là que l'individu s'exprime le plus. Il semble clair que la formation donnée aux futurs enseignants est à davantage adapter aux réalités (article de Colette Corblin et Francine Voltz : « Quel français pensent-ils enseigner ? ») : les stagiaires semblent encore trop nombreux à se trouver démunis devant une norme scolaire très différente des pratiques langagières des élèves. Une étude très rigoureuse de Christopher Stewart et Zsuzanna Fagyal intitulée ici « Engueulade ou énumération ? Attitudes envers quelques énoncés enregistrés dans "les banlieues" » montre qu'entendre certains accents de banlieue provoque effectivement une sensation d'agressivité et de violence… Un enseignant doit apprendre à connaître et à gérer ces effets. C'est donc en terme de rencontres qu'il est nécessaire de penser cet enseignement et l'enseignant doit comprendre qu'il n'est pas face à des jeunes qui forment un groupe homogène : les cultures des jeunes ne peuvent être comprises qu'en tant que mouvements dynamiques. C'est ce que Daniel Delas souligne dans « Professeur de banlieue et culture des jeunes ». La classe peut aussi être le lieu (parfois le seul) qui reconnaît l'histoire individuelle des élèves, et l'on admire le respect constructif qui se dégage de l'expérience de Sabine Contrepois (« Entre mémoire familiale et leçon d'humanité : une interaction pédagogique en lycée professionnel » …). Dans deux articles complémentaires et riches d'enseignement, (« Les élèves de milieux populaires et leurs pratiques langagières face aux évidences et exigences de l'école » et « La maîtrise des discours, un objectif réaliste pour les classes "difficiles" ? ») Elisabeth Bautier et Daniele Manresse soulignent l'écart qui existe entre les pratiques langagières des jeunes et les exigences scolaires. Mais il ne s'agit pas pour elles de s'arrêter à un constat qui prophétiserait l'échec. Elles insistent sur le fait que l'enseignement du français en milieu « difficile » doit voir ici son enjeu majeur : pour que l'école ne mène pas à l'échec systématique, il faut reconnaître que l'élève a un rapport au langage différent de celui

attendu : la différence entre « lui » et les signifiances véhiculées par la langue n'est pas encore acquise, l'individu n'est pas encore sujet de la construction de son savoir scolaire, même s'il l'est probablement dans d'autres contextes. Si l'enseignant se contente d'entendre l'élève sans le faire avancer vers un rapport renouvelé au langage, c'est tout le rapport au savoir qui mène à l'échec scolaire. C'est dans cette dynamique qu'on peut envisager l'article de Marie-Françoise Chanfrault-Duchet (« Enseigner le français dans les banlieues : les enjeux de l'oralité »), celui de Monique Jurado et Alain Merlet (« Enseigner la littérature en banlieue ? Stéréotypes et paradoxes »), ou celui de Jacques David (« l'Ecriture des collégiens de banlieue, entre pratiques singulières et normes scolaires »). Cette réflexion ne peut être évidemment menée que dans le cadre d'une réelle politique d'intégration des nouveaux-arrivants, et l'on voit dans l'article comparatif de Claude Cordier et Mélanie Richet (« Les dispositifs d'accueil et de scolarisation des nouveaux arrivants allophones : un observatoire pour les politiques locales d'intégration/ségrégation ») combien cela dépend des caractéristiques des dispositifs mis en place dans chaque unité d'accueil. Il ne s'agit donc pas de baisser les bras face à la difficulté – réelle – de l'enseignement dans ces banlieues difficiles, mais bien de tenter de mieux en comprendre les particularités, pour une meilleure adaptation.

On souligne dans une courte conclusion combien il est légitime et nécessaire de penser les banlieues comme lieux « d'invention » plutôt que comme milieux « difficiles ».

On appréciera l'approche globale et nouvelle de cet ouvrage qui, de par son principe d'organisation, casse les ghettoï sations et favorise les rencontres interdisciplinaires. Cette organisation, à l'image des espaces dont on parle, permet une lecture-mosaï que, non linéaire, et donne au lecteur cette autonomie et cette liberté dont d'aucuns déplorent le manque dans certaines banlieues. Le risque reste – on le perçoit ici – que la multiplicité des voix gène la synthèse et que l'identité soit cachée par la pluralité, même si le mérite demeure : ne pas confondre identité et unicité. On peut par ailleurs regretter l'absence de certains acteurs privilégiés de ces situations : on pense en particulier aux jeunes eux-mêmes et aux représentants de la loi, ou même à des éducateurs sociaux qui n'agissent pas directement dans l'école.

De façon pragmatique, il est à espérer que la formation des enseignants saura s'inspirer d'une telle démarche, pour une meilleure efficacité d'une école qui peine à gérer ses paradoxes : intégrer et accueillir l'autre dans sa différence, reconnaître la diversité des (futurs) citoyens pour une meilleure reconnaissance de l'identité nationale.

GLOTTOPOL

Revue de sociolinguistique en ligne