GLOTTOPOL

Revue de sociolinguistique en ligne

N° 8 – juillet 2006

Traitements automatisés des corpus spécialisés : contextes et sens

SOMMAIRE

Myriam Mortchev-Bouveret : *Présentation*

Aurélie Névéol et Sylwia Ozdowska : *Terminologie bilingue anglais-français : usages clinique et législatif*

Pierre Zweigenbaum et Benoit Habert : *Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue*

Tran Duc Tuan : *Système de recherche d'information médicale par croisement de langues : vietnamien-français-anglais*

Pierre Beust et Thibault Roy : *Prendre en compte la dimension globale d'un corpus dans la contextualisation du sens : expérimentations en informatique linguistique*

Sylvie Vandaele, Sylvie Boudreau, Leslie Lubin, Elizabeth Marshman : *La conceptualisation métaphorique en biomédecine : indices de conceptualisation et réseaux lexicaux*

Compte rendu

Véronique Miguel : 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 p., ISBN 2-7342-1013-4.

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[1]. 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.

[1] 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[2] 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

[2] Comme indiqué lors 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[3] : 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)[4] consiste à exploiter l'hypothèse distributionnaliste citée plus haut (« deux mots ont

[3] 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.

[4] 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.

un sens proche s'ils sont employés dans des contextes très voisins »), en l'étendant au cas 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[5], 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).

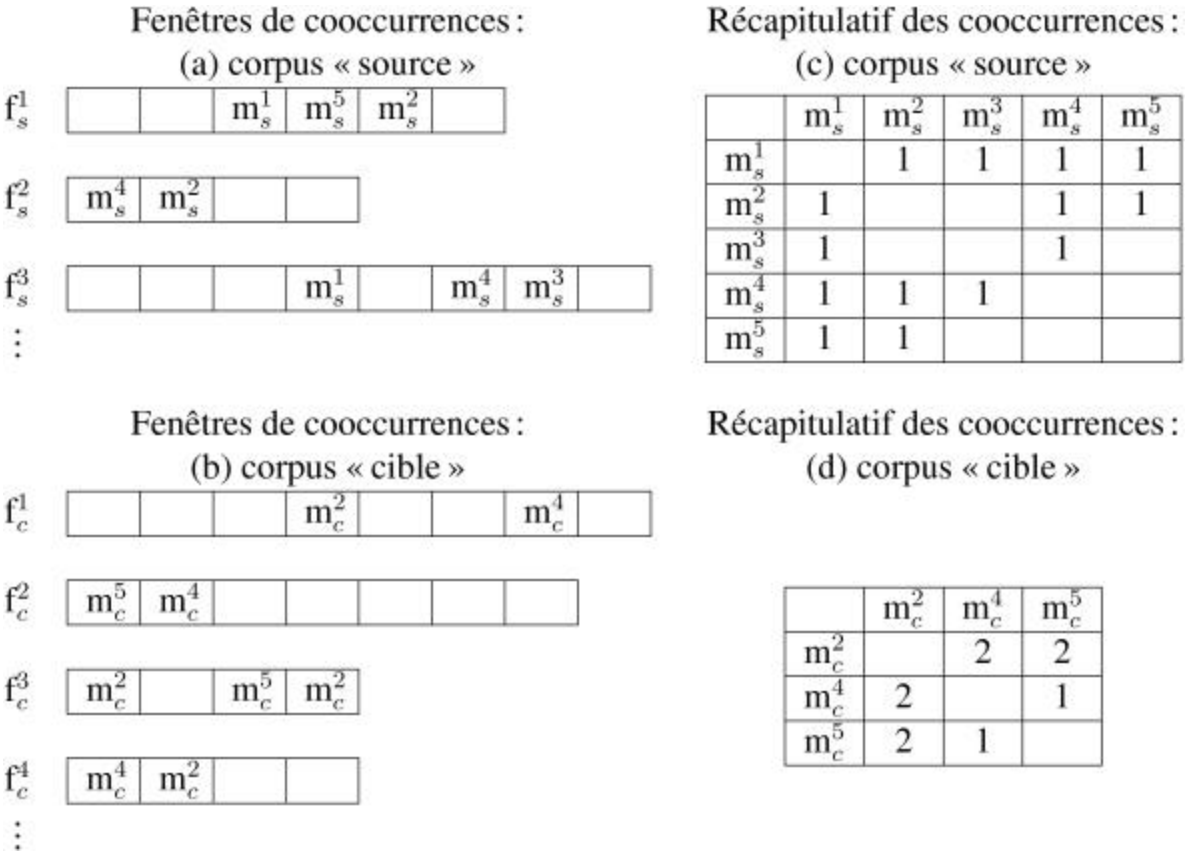

FIG. 2 : Proximités entre mots – corpus comparables (1/2)

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

[5] 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.

cible (ou e^{-1} dans le sens langue cible-langue source) qui permet pour une fenêtre dans la 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[6]. 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.

[6] 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)[7] 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[8] (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[9]. 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.

[7] http://www.chu-rouen.fr/cismef/
[8] http://www.nlm.nih.gov/mesh/meshhome.html
[9] 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[10] 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.

[10] 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[11]. 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

[11] 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)[12]. 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

[12] 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[13] 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[14]. 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

[13] 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.

[14] 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.

GLOTTOPOL

Revue de sociolinguistique en ligne