Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du Web

Stéphanie Léon & Chrystel Millon

Equipe DELIC – Université de Provence
29, Av. Robert Schuman – 13621 Aix-en-Provence Cedex 1
fanny.leon@orange.fr, Chrystel.Millon@up.univ-mrs.fr

Mots-clefs – Keywords

Traduction, corpus, relations lexicales bilingues, acquisition semi-automatique, World Wide Web.

Translation, corpus, bilingual lexical relations, semi-automatic acquisition, World Wide Web.

Résumé – Abstract

Cet article présente une méthode d'acquisition semi-automatique de relations lexicales bilingues (français-anglais) faisant appel à un processus de validation sur le Web. Notre approche consiste d'abord à extraire automatiquement des relations lexicales françaises. Nous générons ensuite leurs traductions potentielles grâce à un dictionnaire électronique. Ces traductions sont enfin automatiquement filtrées à partir de requêtes lancées sur le moteur de recherche Google. Notre évaluation sur 10 mots français très polysémiques montre que le Web permet de constituer ou compléter des bases de données lexicales multilingues, encore trop rares, mais dont l'utilité est pourtant primordiale pour de nombreuses applications, dont la traduction automatique.

This paper presents a method of semi-automatic acquisition of bilingual (French-English) lexical relations using a validation process via the Web. Our approach consists firstly of automatically extracting French lexical relations. We then generate their potential translations by means of an electronic dictionary. These translations are finally automatically filtered using queries on the Google search engine. Our evaluation on 10 very polysemous French words shows that the Web is a useful resource for building or improving multilingual lexical databases, which are urgently needed in a wide range of applications, such as machine translation.

1 Introduction

Bien qu'elle ait été la première application non-numérique de l'informatique, la traduction automatique a connu des débuts décevants, qui ont jeté un discrédit sur cette technologie pendant plusieurs décennies. Toutefois, des progrès ont été accomplis au cours des dernières années, en raison de l'avènement du Web dans un contexte fortement multilingue, de

l'accroissement très important de la couverture des dictionnaires présents dans les systèmes de traduction, et de la prise en compte d'un nombre croissant d'expressions composées. Par exemple, le système Systran[1] traduit désormais correctement du français vers l'anglais des expressions telles que :

vol à main armée → armed robbery
vol à la roulotte → stealing from parked vehicles

En revanche, dès que l'on sort de ces listes d'expressions figées, on retombe rapidement dans des erreurs de traduction qui gênent considérablement la compréhension, et lui donnent même parfois un caractère surréaliste. Ainsi, Systran utilise la traduction la plus fréquente du mot *vol,* c'est-à-dire *flight* (VOL AERIEN), dans toutes les autres situations. Si *réserver un vol* est correctement traduit (*to reserve a flight*), *commettre un vol* est traduit par *to make a flight,* ce qui est totalement incompréhensible dans ce contexte pour un anglophone.

Pourtant, la relation lexicale[2] *commettre-vol* est un indice désambiguïsateur très fort, qui, si elle était correctement traduite dans une base de données (*to commit-theft*), pourrait servir à générer des traductions correctes. La combinatoire est toutefois beaucoup plus ouverte qu'avec les expressions composées mentionnées plus haut et la constitution manuelle d'une base de données de combinaisons lexicales à grande échelle est une tâche à peu près impossible. Les dictionnaires bilingues se contentent d'ailleurs de rares indications ponctuelles, se fiant au jugement du lecteur et à sa connaissance du monde, que l'on ne peut guère espérer d'une machine.

Afin de constituer de telles bases de données bilingues, de nombreux travaux se sont appuyés sur des corpus parallèles ou alignés[3] (Véronis, 2000). Toutefois, ces méthodes présentent diverses limites : de telles ressources sont peu nombreuses et concernent des domaines restreints. Le paysage est un peu plus souriant avec des corpus comparables[4] (Morin, Dufour-Kowalski et Daille, 2004), mais les contraintes restent fortes.

Le Web, qui génère des besoins considérables de traduction, offre en même temps un réservoir gigantesque de données qui peuvent être exploitées par des moyens automatiques, en particulier grâce à des moteurs de recherche tels que Google ou Altavista. En se basant sur les travaux de (Kilgarriff & Grefenstette, 2003), on peut estimer à environ 100 milliards le nombre de mots indexés par Google pour la seule langue anglaise. Même si les données du Web sont moins contrôlées, et donc plus « bruitées », elles permettent d'envisager un changement radical d'échelle pour les méthodes basées sur les données, à condition de développer des méthodes et des techniques adaptées. Depuis quelques années, divers domaines du TAL utilisent le Web en tant que ressource de données linguistique. (Cao & Li, 2002) proposent une méthode automatique de génération et de sélection de traductions pour les syntagmes nominaux de l'anglais vers le chinois à partir du Web. Les résultats montrent que le Web peut être utile tant pour l'acquisition de données que pour une aide à la validation.

---

1 http://www.systransoft.com/

2 Concernant la combinatoire lexicale, la littérature présente une terminologie disparate et souvent floue. Certains parlent de « préférences lexicales » (Wilks, 1975), de « restrictions de sélection » (Katz & Fodor, 1964) ou encore de « collocations » (Benson, 1990 ; Smadja, 1993 ; Cruse, 1986). Afin de désigner ce phénomène, nous employons ici le terme « relation lexicale », plus neutre, défini comme une cooccurrence lexicale entre deux lexèmes liés syntaxiquement.

3 Ensemble de textes alignés avec leur traduction au niveau du paragraphe, de la phrase, des expressions ou des mots.

4 Corpus de langues différentes traitant du même domaine mais non parallèles.

L'objectif de notre article est de proposer une méthodologie de validation semi-automatique de traductions anglaises via le Web, à partir de relations lexicales françaises du type *NOM ADJECTIF*, *NOM1 DE NOM2* et *VERBE NOM*(*objet*) extraites d'un corpus français de pages Web en langue générale. Pour revenir à nos exemples *commettre un vol* et *réserver un vol* (voir Figure 1), Google nous permet de valider les traductions correctes, grâce à leur nombre d'occurrences. Par exemple, la requête [*"commit a flight"* OR *"commit the flight"*] retourne seulement 13 résultats. La requête [*"commit a theft"* OR *"commit the theft"*] retourne quant à elle 5110 résultats. Parmi ces deux traductions candidates, les résultats sélectionnent de façon écrasante la relation lexicale satisfaisante (*to commit-theft*).

	Effectifs absolus		***Effectifs par million***	
	flight	**theft**	**flight**	**theft**
commit	13	5510	0	306
reserve	33 500	3	592	0

Figure 1 : Exemples de résultats sur Google (janvier 2005)

2 Méthodologie et traitement des données

Notre étude comporte trois étapes. Nous procédons d'abord à une acquisition de relations lexicales françaises via un corpus de pages Web. Nous générons ensuite de façon automatique leurs traductions potentielles à partir d'un dictionnaire électronique. Nous interrogeons enfin le moteur de recherche Google afin de valider ces dernières de façon semi-automatique.

2.1 Extraction automatique des relations lexicales françaises

Notre méthode a été testée sur la combinatoire lexicale de 10 noms français très polysémiques (*barrage*, *détention*, *formation*, *lancement*, *organe*, *passage*, *restauration*, *solution*, *station* et *vol)*. Ces mots ont été jugés comme les plus polysémiques parmi 200 noms de fréquence équivalente, lors du projet Senseval (Véronis, 1998) et constituent donc un banc de test difficile, qui a été utilisé par la suite dans divers travaux. Nous exploitons la version lemmatisée et étiquetée morpho-syntaxiquement du corpus de pages Web francophones élaboré par Jean Véronis (Véronis, 2003) autour de ces 10 noms.

L'acquisition automatique des relations lexicales françaises de type *NOM ADJECTIF*, *NOM1 DE NOM2*[1] et *VERBE NOM*(*objet*) utilise une version améliorée du programme employé dans (Millon, 2004), dans lequel des filtres linguistiques d'extraction sont utilisés (représentation de patrons syntaxiques, filtres de candidats indésirables). Les relations lexicales sont ensuite soumises à un seuil limite de fréquence fixé à au moins 10 occurrences, afin d'obtenir des relations lexicales représentatives de besoins en lexicographie. En effet, un nombre important de relations moins fréquentes sont « accidentelles » (comme par exemple *barrage violet*), et non pertinentes pour notre étude. L'objectif est d'extraire une majorité de relations lexicales « propres » du côté français. A l'inverse, des relations lexicales caractéristiques sont perdues (ainsi *barrage hydraulique* a une fréquence de 4).

Une catégorisation sémantique des relations lexicales n'est pas réalisée, car, outre qu'elle serait extrêmement difficile à obtenir avec fiabilité, elle se perdrait lors de l'interrogation des traductions potentielles dans Google. Pour une relation lexicale française telle que *vol de nuit*, l'objectif est de donner un ensemble d'équivalences en anglais qui reflètent soit des variations

[1] Le nom source peut se trouver en position *NOM1* ou *NOM2*.

lexicales pour un même usage comme dans les traductions *night flight* et *night flying* (usage VOL AERIEN), soit des usages différents comme l'exemple de *night robbery* (usage DELIT). La Figure 2 donne la quantité de données obtenues après filtrage des relations lexicales françaises (seuil de fréquence).

	RLs françaises totales	**RLs Françaises ≥ 10**	
		Occurrences	**% restant**
N ADJ	1332	113	8,48%
N DE N	1940	173	8,92%
V N	1278	57	4,46%
TOTAL	**4550**	**286**	**6,29%**

Figure 2 : Résultats de l'extraction des relations lexicales sources

2.2 Génération automatique des traductions potentielles

Les dictionnaires courants contiennent un nombre très restreint de ces relations lexicales, généralement les plus figées. Ainsi, le *Collins Pocket French-English Dictionary* disponible dans l'équipe sous forme électronique grâce à un accord avec l'éditeur Collins, ne propose de traduction que pour 6,6 % des relations lexicales françaises que nous conservons après filtrage, telles que *barrage routier* traduit par *roadblock* ou *station balnéaire* traduite par *seaside resort*.

Pour chacune des relations lexicales françaises non présentes dans le dictionnaire, nous générons automatiquement toutes les traductions possibles via le *Collins Pocket*. Reprenons pour exemple *réserver-vol*. Le *Collins Pocket* donne les traductions suivantes pour les unités lexicales sources *vol* et *réserver* (unité lexicale source vers unité lexicale cible) :

vol → *flight, theft, flying*
réserver → *to reserve, to book*

Notre programme génère toute la combinatoire :

réserver un vol → *to reserve a flight, to reserve a theft, to reserve a flying, to book a flight, to book a theft, to book a flying*

Afin d'avoir un ensemble de traductions le plus exhaustif possible, nous recensons également les « traductions inversées » des unités lexicales françaises, en recherchant ces dernières lorsqu'elles apparaissent en tant que traduction dans la version *English-French*, ce qui rajoute parfois des traductions, comme pour *vol* :

larceny, robbery, snatch → *vol*

Lorsque le dictionnaire propose une traduction (par exemple *vol libre* → *hand-gliding*), nous n'avons pas généré de traduction supplémentaire. Dans certains cas, comme pour *rampe de lancement*, traduite par *launch pad* et *launching pad* dans le dictionnaire, nous manquerons quelques traductions correctes comme *launching ramp* et *launch ramp*.

La Figure 3 donne la quantité de traductions potentielles générées et la proportion de celles-ci par relation lexicale française.

	Traductions générées	Moyenne par RL française
N ADJ	1215	11
N DE N	5155	30
V N	1012	18
TOTAL	7382	19

Figure 3 : Résultats de l'étape de génération des traductions candidates

2.3 Interrogation automatique du moteur de recherche Google

Le moteur de recherche Google a été interrogé automatiquement à l'aide de l'interface de programmation d'applications API (*Application Programming Interface*)[1] afin de récupérer le nombre d'occurrences[2] de chaque relation lexicale anglaise candidate. Ces fréquences seront utilisées lors de la validation[3].

Pour chaque traduction potentielle, nous générons un ensemble de requêtes (voir Figure 4), en considérant les mots de la requête comme une expression exacte, via l'utilisation des guillemets. La recherche est restreinte aux pages Web de langue anglaise.

Patron syntaxique source	Requête (en langue cible)
NOM ADJECTIF	"the ADJ NOM" OR "a ADJ NOM"
VERBE NOM(objet)	"VERBE the NOM" OR "VERBE a NOM"
NOM1 de NOM2	"NOM_1 of NOM_2"
	"NOM_2 NOM_1"

Figure 4 : Patrons des requêtes des relations lexicales anglaises

La combinaison booléenne pour les patrons syntaxiques de type *NOM ADJECTIF* et *VERBE NOM*(*objet*) ramène un ensemble de résultats qui prend en compte les variations dues aux changements d'article, comme dans l'exemple *"commit a theft" OR "commit the theft"*.

L'utilisation d'articles dans les requêtes du patron syntaxique *NOM ADJECTIF* permet également de réduire le problème de l'ambiguïté catégorielle. Par exemple, *complete* peut être un adjectif (*entier, complet, intégral, total*) ou un verbe (*parfaire, compléter*). La relation lexicale *complete restoration* est ambiguë. L'ajout de l'article permet d'éliminer les cas où *complete* est un verbe.

Le patron syntaxique *NOM1 DE NOM2* pouvant être traduit par différentes structures en anglais selon la relation sémantique considérée entre les deux objets, nous traitons séparément deux types de structures dans les requêtes (Chuquet & Paillard, 1987) : d'une part, le patron *N2 N1* marque une relation étroite entre les deux noms et d'autre part, la structure *N1 of N2* accorde la priorité à l'élément repéré (*N2*)[4].

[1] http://www.google.com/apis/

[2] Des différences ont été remarquées entre le nombre de résultats renvoyés par l'API et par l'interface web. Ce problème a été mentionné dans divers forums, mais aucune explication n'a pu être fournie.

[3] Précisons que si les fréquences de Google sont peu fiables dans le cadre de certaines configurations de requêtes dans « tout le Web » (http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html), ce problème ne concerne pas l'utilisation que nous faisons de Google puisque nous limitons les requêtes à une langue donnée.

[4] Le cas du génitif (*N1 's N2*) n'est pas pris en compte dans le cadre de cette étude.

2.4 Validation semi-automatique des traductions potentielles anglaises

2.4.1 Filtre automatique

Afin de réduire le bruit, un filtre simple a été appliqué aux traductions restantes. Nous ne conservons que celles dont la fréquence sur le Web est au moins égale à un millième des occurrences du mot cible. Nous utilisons à l'heure actuelle une méthode statistique simple qui est concluante pour une première approche. Mais nous pouvons envisager par la suite d'autres techniques plus élaborées. Prenons pour exemple *réserver-vol* et deux de ses traductions candidates *book a/the flight* et *book a/the theft* :

Seuil_theft : 2150000 / 1000 = 2150
Seuil_flight : 5760000 / 1000 = 5760

La relation lexicale *book a/the flight* (avec une fréquence de 244000, donc supérieure au seuil limite pour le nom cible *flight*) est retenue, tandis que *book a/the theft* (avec une fréquence de 61, donc inférieure au seuil limite pour le nom cible *theft*) est rejetée.

Ce filtre provoque évidemment parfois des cas de silence. Ainsi, à partir de la relation lexicale *barrage hydro-électrique*, la traduction *hydroelectric dam* est retenue (fréquence de 3920 et seuil à 1910), tandis que *hydroelectric barrage*, également valide, a une fréquence de 32 (pour un seuil à 139) et est rejetée. Notre approche favorise volontairement la précision, car il s'agit de compléter le plus automatiquement possible des ressources existantes. L'augmentation du bruit obligerait à un filtrage manuel des résultats beaucoup plus long et coûteux. Après le filtre automatique sur les fréquences, 7,5 % des relations anglaises potentielles sont conservées (Figure 5).

2.4.2 Validation manuelle

Une validation manuelle nous permet d'évaluer les relations lexicales restantes après filtre automatique, en faisant appel à divers dictionnaires de langue, ainsi qu'à nos connaissances. Les traductions candidates les plus « délicates » sont vérifiées en contexte sur le Web, par reformulation de la requête à travers Google, ainsi que soumises au jugement de plusieurs locuteurs. Nous détaillons les résultats dans la section suivante.

	Traductions générées	***Filtre automatique***		***Validation manuelle***	
		Filtre seuil fréquence		**Traductions valides**	
N ADJ	1215	136	11,2%	132	97,1%
N DE N	5155	351	6,8%	270	76,9%
V N	1012	63	6,2%	56	88,9%
TOTAL	**7382**	**550**	**7,5%**	**458**	**83,3%**

Figure 5 : Résultats de la validation des traductions

3 Premiers résultats

La précision globale des traductions extraites est de 83,3%. La méthode fonctionne particulièrement bien pour les patrons syntaxiques *NOM ADJECTIF* (97,1%) et *VERBE NOM* (*objet*) (88,9%) (Figure 6). Le patron *NOM1 DE NOM2* pose davantage de difficultés en traduction (76,9%).

MOT SOURCE	N ADJ		N de N		V N		TOTAL	
	Nb total	*% valide*	*Nb total*	*% valide*	*Nb total*	*% valide*	*Nb total*	*% valide*
Barrage	39	94,9	33	81,8	19	94,7	91	90,1
Détention	15	100,0	114	80,7	5	100,0	134	83,6
Formation	16	100,0	76	72,4	4	75,0	96	77,1
Lancement	7	100,0	17	88,2	2	100,0	26	92,3
Organe	27	96,3	21	76,2	4	25,0	52	82,7
Passage	10	90,0	21	57,1	13	92,3	44	75,0
restauration	2	100,0	23	78,3	Pas de RL	Pas de RL	25	80,0
Solution	14	100,0	5	100,0	8	100,0	27	100,0
Station	4	100,0	18	72,2	4	100,0	26	80,8
Vol	2	100,0	23	73,9	4	75,0	29	75,9
TOTAL	**136**	**97,1**	**351**	**76,9**	**63**	**88,9**	**550**	**83,3**

Figure 6 : Précision globale des relations lexicales anglaises évaluées

La Figure 7 présente le nombre moyen de traductions valides par relation lexicale française. En moyenne, on obtient deux traductions correctes pour chaque relation lexicale française.

	RLs françaises	**Traductions validées**	**Moyenne par RL**
N ADJ	63	132	2,1
N de N	124	270	2,2
V N	31	56	1,8
MOYENNE	**218**	**458**	**2,1**

Figure 7 : Nombre moyen de traductions valides par relation lexicale française

La Figure 8 présente un exemple de traductions obtenues pour les relations lexicales *barrage hydro-électrique*, *construction de barrage* et *construire-barrage*.

PATRON	**RL FRANCAISE**	**TRADUCTION**
N ADJ	barrage hydro-électrique	hydroelectric dam
N de N	construction de barrage	barrage building
		barrage construction
		barricade building
		barricade construction
		dam building
		dam construction
		weir building
		weir construction
V N	construire-barrage	to build-barrage
		to build-barricade
		to build-dam
		to build-roadblock
		to construct-dam
		to erect-barricade
		to erect-roadblock

Figure 8 : Traductions des relations lexicales de barrage

Les traductions obtenues permettent de multiplier par 10 la quantité de relations lexicales déjà présentes en entrée dans le dictionnaire pour le patron *NOM ADJECTIF*, par 45 celles pour le patron *NOM1 DE NOM2* et par 56 pour le patron *VERBE NOM*(*objet*). L'apport de la méthode est donc appréciable.

4 Analyse des problèmes et perspectives

4.1 Erreurs et difficultés

Les erreurs et les limites de notre stratégie sont catégorisées selon les types de problème et les perspectives d'amélioration.

4.1.1 Erreurs syntaxiques

Certaines erreurs sont dues à des problèmes d'ambiguïté de rattachement syntaxique concernant les traductions candidates testées sur l'API Google. Le moteur de recherche ne permettant pas un accès direct aux catégories morpho-syntaxiques, une analyse syntaxique des contextes des traductions candidates n'a pu être réalisée. Prenons l'exemple suivant :

*The Library of Congress **set the changeover** date.*

La relation lexicale recherchée est *set the changeover*. Or dans ce cas, *changeover* est régi par le nom *date* et non par la forme verbale *set*. Une autre limite concerne les problèmes d'ambiguïté catégorielle. Les formes en *-ing*, par exemple, sont propices à ce type d'erreurs. Une perspective d'amélioration des problèmes d'ordre syntaxique est donc manifestement l'application d'un analyseur syntaxique aux pages Web anglaises.

Enfin, il arrive qu'une relation lexicale ne soit pas traduite par une structure de même longueur. Par exemple, *barrage routier* se traduit par une unité lexicale simple : *roadblock*. La formation des mots composés en anglais, avec ou sans espace ou tiret, est évidemment un cas extrêmement difficile, mais on peut envisager de générer des requêtes du type N-N ou NN (sans espace).

4.1.2 Erreurs sémantiques

Un type d'erreurs d'ordre sémantique concerne l'acquisition de relations lexicales anglaises valides, mais non correspondantes à la relation lexicale source, comme dans l'exemple :

cours de formation --> group rate (59900 occurrences)

Ici, *group rate* signifie *tarif de groupe*.

De plus, la traduction d'une relation lexicale n'est pas toujours obtenue de façon compositionnelle (Melamed, 2001, cité par Morin *et al*., 2004). Par exemple, en anglais, il n'existe pas une traduction littérale de *forcer un barrage* : la traduction va dépendre du contexte situationnel (*to drive through a roadblock*, *to run through a roadblock*, etc.*)*.

Une autre limite est due à l'absence de certains usages dans le dictionnaire. C'est le cas par exemple de l'acception sportive du nom *barrage* (*playoff*). Ainsi, la relation lexicale *match de barrage* est traduite par *weir game* et *barrage game* au lieu de la traduction correcte *playoff game*.

4.1.3 Erreurs techniques

Google ne prend pas en compte des phénomènes tels que la ponctuation, les majuscules, ou la marque du génitif (*'s*), lors des recherches sur le Web. Certaines relations lexicales anglaises

erronées, comme *to reserve-theft* (fréquence de 3) n'ont pas une fréquence nulle pour cette raison. L'exemple suivant montre que les mots *reserve* et *a theft* appartiennent à deux syntagmes différents :

> *A man will face court next month charged with stealing three date palms from a Swansea* ***reserve, a theft*** *which sparked three months of community outrage.*

Le filtre sur les fréquences permet d'éliminer une partie des relations lexicales erronées. Néanmoins, ces problèmes concernent également des traductions correctes, et « bruitent » quelque peu les fréquences de l'API Google[1].

4.2 Perspectives d'amélioration du protocole

4.2.1 Changement de seuil de filtrage des relations lexicales françaises

Les relations lexicales françaises qui comptent moins de 10 occurrences au sein du corpus sont éliminées. Or, un certain nombre de relations lexicales correctes ont des fréquences inférieures à ce seuil. C'est le cas, par exemple pour l'usage sportif de *barrage* (*match de barrage, barrage aller, barrage retour,* etc.). L'algorithme *HyperLex* (Véronis, 2003, 2004) nous permettrait d'identifier les usages peu fréquents des mots (jusqu'à environ 1% des occurrences). Une amélioration consisterait à ajuster le seuil des relations lexicales françaises selon la fréquence de l'usage du nom concerné.

4.2.2 Description des patrons syntaxiques de l'anglais

Contrairement à notre méthode d'extraction des relations lexicales françaises, ne sont pris en compte que les patrons de « base » de l'anglais sans autres variations que celles de l'article dans le cadre de notre acquisition des traductions. Ces patrons « de base » ont donné un premier éventail de résultats qui ont permis d'évaluer notre méthode. Des améliorations ultérieures visent à augmenter le panel de patrons morpho-syntaxiques anglais des traductions candidates, en procédant à une analyse syntaxique du contenu des pages Web prospectées.

4.2.3 Variations morpho-syntaxiques au sein des requêtes anglaises

Une amélioration vise à considérer des variations morpho-syntaxiques au sein de nos requêtes telles que des variations dues aux formes verbales ou au pluriel, comme dans l'exemple :

> *"commit a theft", "commit the theft", "commit thefts" , "commit the thefts", "committed a theft",* etc.

5 Conclusion

Nous avons décrit une méthode d'acquisition semi-automatique de relations de traductions du français vers l'anglais, en montrant que le Web s'avère être un outil particulièrement efficace d'aide à la validation de traductions candidates. Les résultats sont particulièrement intéressants pour les patrons syntaxiques de type *NOM ADJECTIF* (précision de 97,1 %) et *VERBE NOM*(*objet*) (précision de 88,9 %). La méthode reste imparfaite pour le patron *NOM1 DE NOM2*, mais le taux de précision est honorable (76,9%), surtout étant donné la difficulté

[1] Une autre limite de l'API Google est de ne pouvoir lancer que 1000 requêtes par jour, ce qui impose des contraintes de temps. Depuis, mars 2005, l'API Yahoo a été lancé et offre une possibilité de 5000 requêtes par jour.

volontaire du banc de test choisi (mots très polysémiques). Même en l'état, la méthode permet un accroissement important des relations lexicales contenues dans notre base de données lexicale. Ces résultats constituent un premier échantillon significatif des possibilités qu'offre le Web en matière de validation de relations lexicales. Nos perspectives d'évolution concernent principalement la mise en place d'une analyse syntaxique des relations candidates en contexte, ainsi qu'une désambiguïsation des relations lexicales à traduire.

Références

Benson M. (1990), Collocations and general-purpose dictionaries, *International Journal of Lexicography*, vol. 3(1), pp. 23-35.

Cao Y., Li H. (2002), Base noun phrase translation using web data and the EM algorithm. In *Proceedings of CoLing*.

Chuquet H., Paillard M. (1987), *Approche linguistique des problèmes de traduction: anglais-français*, Gap, Ophrys.

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

Katz J. J, Fodor J. A. (1964), The structure of a semantic theory, In J. A. Fodor and J. J. Katz, editors, *The Structure of Language,* chapter 19, pp. 479-518.

Kilgarriff A., Grefenstette G. (2003), Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics, 29*(3): 333-348.

Melamed I. D. (2001), *Empirical methods for exploiting parallel texts*, MIT Press.

Millon C. (2004), Acquisition de relations lexicales désambiguïsées à partir du Web, *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2004),* Fès, (Maroc).

Morin E., Dufour-Kowalski S., Daille B. (2004), Extraction de terminologies bilingues à partir de corpus , *Actes de TALN'2004*, Fès (Maroc).

Smadja F. (1993), Retrieving collocations from text : Xtract, *Computational Linguistics*, Vol. 19, pp. 143-177.

Véronis J. (1998), A study of polysemy judgements and inter-annotator agreement, *Programme and advanced papers of the Senseval workshop,* pp. 2-4, Herstmonceux Castle (England).

Véronis J. (Ed.) (2000). Parallel Text Processing: Alignment and use of translation corpora, Kluwer Academic Publishers.

Véronis J. (2003), Hyperlex : cartographie lexicale pour la recherche d'informations, *Actes de TALN'2003,* pp. 265-274, Batz-sur-mer (France): ATALA.

Véronis J. (2004), *HyperLex : cartographie lexicale pour la recherche d'informations.*Rapport Interne Equipe DELIC, Université de Provence. [En ligne : http://www.up.univmrs.fr/veronis/pdf/2004-hyperlex-rapport.pdf]

Wilks Y. A. (1975), Preference Semantics, In: Keenan, E. (ed), *The Formal Semantics of Natural Language,* Cambridge University Press, pp. 329-348.