Traitement de la polysémie lexicale dans un but de traduction

Marianna Apidianaki

Traitement de la polysémie lexicale dans un but de traduction

Marianna Apidianaki

Lattice - CNRS / ENS / Université Paris 7 - Denis Diderot
Ecole Normale Supérieure, 1, rue Maurice Arnoux, F-92120 Montrouge
Marianna.Apidianaki@linguist.jussieu.fr

Résumé La désambiguïsation lexicale a une place centrale dans les applications de Traitement Automatique des Langues relatives à la traduction. Le travail présenté ici fait partie d'une étude sur les recouvrements et les divergences entre les espaces sémantiques occupés par des unités polysémiques de deux langues. Les correspondances entre ces unités sont rarement biunivoques et l'étude de ces correspondances aide à tirer des conclusions sur les possibilités et les limites d'utilisation d'une autre langue pour la désambiguïsation des unités d'une langue source. Le but de ce travail est l'établissement de correspondances d'une granularité *optimale* entre les unités de deux langues entretenant des relations de traduction. Ces correspondances seraient utilisables pour la prédiction des équivalents de traduction les plus adéquats de nouvelles occurrences des éléments polysémiques.

Abstract Word Sense Disambiguation has a central role in NLP applications relevant to translation. The work presented in this article is a part of a study on the overlaps and divergences existing between the semantic spaces occupied by the polysemous items of two languages. Correspondences between those items are rarely biunivocal and their study gives us insights on the possibilities and limits of using a second language for the disambiguation of polysemous items of a source language. The aim of this work is the establishment of correspondences of an *optimal* granularity between the items of two languages being in relation of translation that could be used for the prediction of the most adequate translation equivalents for new occurrences of the polysemous source language items.

Mots-clés: polysémie, cooccurrences, correspondances de traduction, prédiction de traduction

Keywords: polysemy, cooccurrences, translation correspondences, translation prediction

1 La polysémie dans un contexte bilingue

La notion du *sens* peut être appréhendée de différentes manières et la diversité des représentations sémantiques existantes démontre la difficulté de trouver un consensus concernant la nature du sens et sa description optimale dans le cadre du traitement automatique. Une méthode de désambiguïsation sémantique assez largement utilisée consiste en l'utilisation de traductions en tant que ressource pour le repérage et la distinction des sens des unités polysémiques d'une langue. On retrouve cette idée dans des travaux importants visant non seulement la désambiguïsation mais aussi l'annotation sémantique et la sélection d'équivalents de traduction pour la Traduction Automatique (Gale et al., 1992, 1993 ; Dagan et al., 1991 ; Teubert, 2002 ; Dyvik, 2003).

Les langues divisent leur espace sémantique de manières variées et même les *régions* sémantiques occupées par des unités lexicales considérées comme équivalentes du point de vue de la traduction peuvent être très différentes. Altenberg et Granger (2002) caractérisent les cas où les unités de deux langues qui se trouvent en relation de traduction ont des extensions de sens différentes comme des cas de *polysémie divergente* (*divergent polysemy*) en les distinguant des cas de polysémie avec *recouvrement de sens* (*overlapping polysemy*), où les unités ont approximativement les mêmes extensions de sens. D'autre part, Salkie (1997) et Viberg (2002) soulignent que les équivalents de traduction décrits dans les dictionnaires ont rarement la même distribution dans des textes réels et que leur degré de correspondance mutuelle est très bas.

Dans les cas de polysémie avec *recouvrement de sens*, la polysémie lexicale est préservée dans la langue cible (LC). Ceci signifie que les unités polysémiques de la langue source (LS) sont traduites par des unités polysémiques équivalentes dans la LC. D'après Salkie (2002), ces unités ne sont pas considérées polysémiques ou ambiguës du point de vue de la traduction et les équivalents n'indiquent pas des distinctions de sens dans la LS. De tels cas sont souvent observés parmi des langues proches. Dans ces cas, le traducteur (ou la machine) n'a pas besoin de résoudre la polysémie lexicale pour traduire. Si la désambiguïsation de ces unités était souhaitée, il faudrait recourir à une troisième (ou quatrième…) langue.

Dans cet article, nous allons proposer une méthode d'exploration des relations existantes entre les unités polysémiques de deux langues. La présentation de cette méthode sera illustrée par des exemples extraits d'un corpus parallèle sur lequel nos expériences ont été menées.

2 Prétraitement du corpus

Le corpus utilisé dans ce travail est un corpus parallèle anglais-grec de 4 000 000 de mots, lemmatisé, morpho-syntaxiquement étiqueté et aligné au niveau des phrases (Gavrilidou *et al.,* 2004). L'alignement a été effectué par l'outil d'alignement de phrases intégré dans le système de Mémoire de Traduction TrAid (Triantafyllou et al., 2000) et validé à la main. La source des textes est le *Journal de l'Union Européenne*. Ces textes relèvent de cinq domaines : droit, éducation, environnement, santé et tourisme. Le corpus a été traité par l'outil Syntex et les résultats sont consultables via l'interface TermOnto (Bourigault et al., 2004). Des sous-corpus sont créés correspondant aux mots polysémiques et contenant les contextes lexicaux (ou co-textes) dans lesquels ils apparaissent. La taille du contexte coïncide avec la taille des segments de traduction, qui contiennent de 0 à 2 phrases par langue. Par exemple, un alignement *2:1* met en correspondance 2 phrases du texte de la LS avec 1 phrase du texte

de la LC, à l'intérieur d'un segment, et un alignement *1:0* indique qu'une phrase du texte de la LS n'a pas de correspondance dans le texte de la LC. La mise en correspondance de 2 phrases de la LS et de 2 phrases de la LC permet de capter les correspondances croisées, c'est-à-dire les cas où l'ordre de 2 phrases dans le texte de la LS est renversé dans la LC. Le choix des segments de traduction en tant que contexte est dicté par notre objectif d'exploration de l'influence du co-texte proche des unités de la LS sur la sélection des équivalents de traduction.

A partir des contextes de chaque mot polysémique on construit une liste de fréquence. Pour le moment, nous nous sommes intéressés seulement aux mots de certaines catégories grammaticales : les noms et les adjectifs. Les noms et les adjectifs qui cooccurrent dans les segments de traduction avec le mot polysémique que nous voulons désambiguïser sont d'abord ramenés à leur lemme. Ensuite on construit la liste des fréquences cumulées associées aux lemmes. La fréquence cumulée d'un lemme correspond à la somme des fréquences des occurrences (ou formes fléchies) associées à ce lemme et trouvées dans les contextes. Par exemple, si dans les textes on a les occurrences *teacher* et *teachers* avec les fréquences *5* et *6* respectivement, la fréquence du lemme *teacher* dans la liste sera *11*.

Les fréquences utilisées dans nos calculs sont les fréquences des formes (*types*) auxquelles les occurrences (*tokens*) sont ramenées. Le calcul basé sur les occurrences peut introduire dans le réseau de cooccurrences sémantiques des liens faux, qui peuvent être dus à des idiomes particulaires et non à la similarité sémantique *authentique*, comme il est souligné dans (Widdows, Dorow, 2002). La lemmatisation a aussi une grande importance pour les langues à morphologie riche, comme le grec. Il faut ajouter que l'ordre syntaxique et les informations positionnelles des mots ne sont pas pris en compte.

3 La désambiguïsation dans un contexte monolingue

Les principes théoriques sous-jacents à ce travail sont ceux de l'approche contextuelle du sens (Firth, 1957 : 11). Suivant cette approche, le sens des mots correspond à leur usage dans les textes. Ainsi, pour les éléments polysémiques on fait l'hypothèse que leurs sens peuvent être distingués en regroupant leurs usages et que leurs contextes peuvent servir à distinguer leurs usages.

Dans un premier temps, on procède à la désambiguïsation des unités polysémiques de la LS à l'aide d'une méthode de cooccurrence proche de celle présentée dans (Véronis, 2003). Les listes de fréquence construites pendant le prétraitement du corpus seront utilisées ici. De la liste de fréquence correspondante à un mot polysémique on ne retient que les formes qui ont une fréquence supérieure à 2. Tout d'abord on choisit le mot avec la plus grande fréquence dans la liste – par exemple, le mot *school* pour le mot polysémique *class*. On considère l'ensemble des contextes où ce mot apparaît et on en construit une matrice de cooccurrence. Les éléments de la matrice sont les lemmes auxquels sont ramenés les noms et les adjectifs qui cooccurrent avec *school* dans les textes. Chaque case [*i,j*] de la matrice de cooccurrence construite pour *school* contient la fréquence de cooccurrence des formes fléchies associées au lemme *i* et des formes fléchies associées au lemme *j* dans l'ensemble des contextes du mot *school*. De cette matrice on ne retient que les cooccurrents qui ont une fréquence supérieure à 1. Les seuils de fréquence et de cooccurrence sont bas à cause de la petite taille de notre corpus.

A partir de la liste de fréquence du mot polysémique et de la matrice de cooccurrence du mot *school* on construit un graphe correspondant à ce dernier. Les nœuds de ce graphe sont les mots retenus après le filtrage des contextes où *school* apparaît. Les arcs reliant les nœuds sont pondérés par la formule suivante :

$$w_{A,B} = 1 - \max\left[p\left(A|B\right), p\left(B|A\right)\right]$$

où $p\left(A|B\right)$ est la probabilité conditionnelle d'observer l'élément A dans un contexte où l'élément B apparaît et inversement pour $p\left(B|A\right)$. Cette mesure reflète la *distance* sémantique entre les mots ; quand elle vaut 0, les mots sont toujours associés tandis que quand elle vaut 1, ils ne le sont jamais. Les arcs qui ont un poids >0.9 sont éliminés. Ainsi les arcs dans le graphe final relient les nœuds qui se trouvent en relation de cooccurrence *significative*.

Une fois le premier graphe construit, on élimine de la liste de fréquence du mot polysémique le mot *school* ainsi que tous ses voisins dans le graphe. De la liste qui en résulte on choisit le mot le plus fréquent – dans notre exemple c'est le mot *number* – et on continue de la même manière. On retient l'ensemble des contextes où *number* apparaît dans le corpus du mot polysémique *class*, on construit la matrice de cooccurrence, on crée le graphe correspondant et on élimine de la liste de fréquence le mot *number* ainsi que ses voisins. On itère ce processus tant que le mot le plus fréquent dans la liste de fréquence a au moins 6 voisins propres.

L'hypothèse sous-jacente à ce processus est que les sens-usages différents du mot polysémique que l'on veut désambiguïser sont décrits par les petits graphes qui lui correspondent. L'utilisation de graphes différents n'exclut pas qu'il peut y avoir des liens entre les sens différents. Ceux-ci peuvent être facilement trouvés à partir des contextes. Le tableau 1 contient des informations quantitatives sur les mots polysémiques étudiés. Chaque sens est décrit à l'aide du mot le plus fréquent qui déclenche la construction du graphe correspondant. Le mot le plus fréquent est aussi le nœud avec le plus grand degré dans le graphe.

Mots polysémiques	Equivalents en grec [1]	Corpus d'entraînement[2]	Corpus de test	Sens-usages	Nombre de nœuds	Nombre d'arcs	Densité du graphe
class	τάξη(201), κατηγορία(20), μάθημα(9), τμήμα(3), εκπαίδευση(2), σώμα(1)	236	59	school	125	467	0.06
				number	23	30	0.11
				device	17	40	0.29
				lesson	11	21	0.38
competence	αρμοδιότητα(118), ικανότητα(88), επάρκεια(4), δικαιοδοσία(4),δεξιότητα(3) δυνατότητα(2), κύρος(1)	220	55	member	143	617	0.06
				skill	29	54	0.13
				qualification	12	19	0.28
movement	κυκλοφορία(251), διακίνηση(38), κίνηση(28), μετακίνηση(19), κίνημα(11), κινητικότητα(6)	353	88	free	245	2177	0.07
				freedom	95	447	0.1
				relation	23	113	0.44
				restriction	14	23	0.25

Tableau 1 : Informations quantitatives sur les mots polysémiques étudiés

4 Correspondances entre sens et équivalents

Pour chacune des unités polysémiques anglaises retenues, on procède au repérage de leurs équivalents en grec. Les correspondances de traduction sont relevées à la main. Cette phase

[1] Ordonnés en fonction de leur fréquence dans le corpus d'entraînement.
[2] Décrit en nombre de segments. De même pour le corpus de test.

pourrait être automatisée en utilisant un outil d'alignement de mots ou un outil de repérage de traduction (Véronis, Langlais, 2000 ; Simard, 2003). Pour le moment, on a choisi le repérage manuel afin d'étudier minutieusement les relations entre les unités des deux langues, de repérer même les équivalents très rares et d'étudier des cas comme les omissions, les ajouts et les reformulations dans la LC.

Une fois les graphes de la LS construits et les équivalents grecs trouvés, la prochaine étape consiste à mettre en relation les graphes de la LS avec les équivalents de traduction repérés. Pour cela, on utilise le contexte des équivalents. Le contexte d'un équivalent est composé du contexte lexical de l'unité de la LS (dans le même segment de traduction) quand elle est traduite par cet équivalent précis. On construit alors, au sein de notre corpus correspondant au mot polysémique, des ensembles de segments correspondant à chacun des équivalents. Pour chaque équivalent, on procède au même calcul de cooccurrents que dans la LS. D'abord, on construit une matrice de cooccurrence à partir des segments qui lui correspondent. Ensuite, en utilisant cette matrice et la liste de fréquence du mot polysémique, on calcule le poids $w_{A,B}$ entre les cooccurrents dans les segments retenus. Ainsi, les mots qui ont une relation de cooccurrence *significative* avec l'équivalent sont-ils retenus. De cette manière, on arrive à construire pour chacun des équivalents de la LC un graphe de cooccurrences qui comprend ses cooccurrents dans la LS.

A partir des graphes de la LS et de la LC, on peut établir des correspondances entre les équivalents de traduction et les sens repérés dans la LS en estimant leur similarité en termes de partage de traits (Tversky, 1977). Pour cela, on calcule le taux de *recouvrement* entre le graphe correspondant à chaque équivalent et les graphes construits pour le mot polysémique. Les traits sur lesquels porte ce calcul ne sont pas les cooccurrents individuels des équivalents et du mot polysémique – qui correspondent aux nœuds du graphe – mais les paires de nœuds qui représentent des relations significatives entre eux. Si, par exemple, dans le graphe d'un équivalent, nous avons les associations *school--timetable* et *primary--education*, nous allons d'abord chercher dans le graphe de la LS les nœuds correspondant aux mots *school*, *timetable* etc., et s'ils existent nous allons voir s'il y a des arcs qui les relient ; si ce n'est pas le cas, on ne les retient pas. Le calcul de recouvrement entre les cooccurrents individuels introduit de faux liens entre les équivalents et les composantes, à cause de la polysémie qui est assez fréquente parmi les cooccurrents individuels. Par contre, la prise en compte des relations que ces éléments entretiennent avec leurs cooccurrents restreint leur ambiguïté. Les liens proposés de cette manière sont beaucoup plus pertinents. L'algorithme utilisé pour calculer le recouvrement entre, d'une part, le graphe d'un équivalent et, d'autre part, un graphe de la LS est décrit ci-dessous :

```
recouvrement (G_LC, G_LS){
        G_LC : graphe de l'équivalent
        G_LS: un des graphes du mot polysémique
        E_LC ← liste des arcs dans G_LC
        E_LS ← liste des arcs dans G_LS
        I ←  Ø
        foreach e1 ∈ E_LC{
                foreach e2 ∈ E_LS {
                        if (e1 = e2)
                        then {
                                I ← e1
                        }
                }
        }
        return I
}
```

Dans le tableau 2, on décrit les correspondances entre les sens-usages des mots polysémiques et leurs équivalents en grec. Dans la deuxième colonne, les sens-usages d'un mot polysémique sont décrits par le mot le plus fréquent dans le graphe correspondant. La troisième colonne contient quelques voisins directs de ce mot pour illustrer ces différents usages. Par exemple, les voisins du mot *device* qui décrit un usage de *class* montrent qu'il est question dans les textes de *classes d'appareils médicaux*.

Dans la colonne droite du tableau, on trouve les équivalents grecs correspondant à chaque sens comme cela a été démontré par le calcul de recouvrement. On remarque que l'on peut avoir plusieurs équivalents mis en relation avec un sens, ou, au contraire, un équivalent mis en relation avec plusieurs sens. Ceci s'explique ainsi : le calcul de recouvrement établit souvent des correspondances entre des sous-ensembles d'un graphe de la LS et des graphes correspondant à des équivalents différents ou entre des sous-ensembles du graphe d'un équivalent et des graphes de sens différents. La correspondance entre un sens et un équivalent est décrite par l'ensemble contenant les traits qui leur sont communs.

Mot polysémique	Sens-usages	Voisins	Equivalents en grec
class	school	classroom, teacher, pupil, elementary…	τάξη, μάθημα, εκπαίδευση, τμήμα
	number	minimum, maximum, total, high, average…	τάξη, κατηγορία
	device	implant, breast, instruction, practice, medical…	κατηγορία
	lesson	orientation, written, second, basis…	τάξη
competence	member	state, sphere, infringement, power, exercise…	αρμοδιότητα, δικαιοδοσία, ικανότητα, κύρος, επάρκεια
	skill	personal, lifelong, language, mathematics…	ικανότητα, επάρκεια
	qualification	recognition, development, partner, trust…	αρμοδιότητα, κατάρτιση
movement	free	student, goods, worker, barrier, citizen…	κυκλοφορία, μετακίνηση, κίνηση, κινητικότητα, διακίνηση, κίνημα
	freedom	border, residence, territory, immigration…	μετακίνηση, διακίνηση, κινητικότητα,κυκλοφορία, κίνηση
	relation	Sweden, Finland, Belgium, Italy, Spain…	κίνηση, κυκλοφορία, κινητικότητα
	restriction	animal, disease, trade, health, risk...	κίνηση, διακίνηση

Tableau 2 : Correspondances entre sens et équivalents des mots polysémiques

Une autre remarque qui s'impose : les *sens* détectés seraient plus facilement caractérisés comme *usages*. Ceci constitue une des caractéristiques des méthodes de désambiguïsation basées sur les cooccurrences. Pour arriver à des *sens*, il faut souvent fusionner des usages qui ont été détectés à partir du corpus. Il ne faut pas aussi sous-estimer le rôle de la nature des textes sur le repérage de sens. Le corpus utilisé dans cette étude est constitué par des textes communautaires, ce qui induit l'utilisation d'une terminologie plus ou moins uniforme, même si ces textes relèvent de sous-domaines différents. Cette donnée a été conservée, car nous souhaitions appliquer notre méthode en utilisant un minimum de connaissances linguistiques. Il reste à explorer si la répétitivité observée dans les textes déguise des distinctions sémantiques plus pertinentes. Néanmoins, les usages proposés peuvent être utilisés pour une première tentative de validation de nos hypothèses.

5 Prédiction de traduction

Dans le paragraphe 4, on a décrit comment les correspondances entre sens et équivalents sont établies. Ces correspondances – ou, plus concrètement, les ensembles de traits qui les décrivent – peuvent être utilisées dans le processus de prédiction des équivalents de traduction les plus adéquats pour de nouvelles occurrences des mots polysémiques. Si l'on avait établi

des correspondances au niveau des mots, celles-ci seraient très *grossières* ; *competence*, par exemple, serait mis en relation avec 7 équivalents en grec. Le calcul de recouvrement entre sens et équivalents permet de restreindre les choix parmi les équivalents, à l'aide de correspondances de granularité plus *fine*.

L'utilisation des correspondances établies pendant la phase précédente permet, d'une part, de désambiguïser la nouvelle occurrence du mot polysémique dans la LS et, d'autre part, de la traduire correctement dans la LC. Dans le cas d'une nouvelle occurrence d'un mot polysémique, c'est le contexte lexical du mot qui va nous guider dans le choix de l'équivalent le plus correct. De ce contexte on garde les noms et les adjectifs et l'on construit une liste contenant des associations entre eux, qui montrent leur relation de cooccurrence dans ce segment de texte. Si l'on a, par exemple, la phrase d'entrée « *The resource teacher prepares materials, which the class teacher can use if necessary* », on construit les associations: *resource -- teacher*, *resource -- material*, *class -- material*, *necessary -- material*, *class – teacher*, etc. Cet ensemble d'associations décrit le contexte de la nouvelle occurrence.

L'ensemble d'associations construit est par la suite comparé aux résultats de l'étape précédente, c'est-à-dire aux ensembles de traits qui décrivent les correspondances entre les sens et les équivalents et qui ont été constitués à partir du corpus d'entraînement. Des ensembles précédemment établies on ne retient que celui/ceux qui partage/partagent des traits avec l'ensemble construit à partir du nouveau contexte. Il s'agit alors, encore une fois, d'un calcul de similarité en termes de partage de traits.

Pour la phrase d'entrée citée plus haut, l'ensemble qui a été retenu est celui qui décrit la correspondance entre le sens *school* du mot polysémique *class* et l'équivalent *τάξη*. De cette manière, on trouve, d'une part, le sens avec lequel le mot polysémique de la LS est utilisé dans ce nouveau contexte (*school*) et, d'autre part, la traduction la plus adéquate pour cette nouvelle occurrence (*τάξη*). Il arrive parfois que l'on trouve des relations entre le contexte de la nouvelle occurrence et plusieurs des ensembles préétablis. Ces relations ne sont pas très nombreuses (de 2 jusqu'à 4). Alors, même si l'on n'arrive pas à avoir une proposition de traduction unique pour la nouvelle occurrence, on arrive à restreindre les choix de traduction. Dans ces cas, il est aussi possible d'attribuer une *préférence* à un équivalent et à un sens en fonction de la quantité et des poids des associations communes, en n'excluant pas les autres.

Néanmoins, si notre but réside seulement dans la prédiction de l'équivalent de traduction correct – et non pas la désambiguïsation de la nouvelle occurrence – on peut utiliser uniquement les graphes correspondant aux équivalents. Les ensembles d'associations des graphes sont comparés à l'ensemble d'associations construit à partir du nouveau contexte et celui/ceux qui partagent le plus de traits avec celui-ci est/sont retenu(s).

6 Evaluation

Afin d'évaluer les méthodes de prédiction décrites dans le paragraphe 5, nous avons utilisé une partie (20 %) du corpus correspondant à chaque mot polysémique, que nous avions mise à part dès le début. La fréquence d'utilisation des équivalents de traduction dans notre corpus étant très variable, nous avons essayé d'inclure dans le corpus d'évaluation des segments correspondant à tous les équivalents. Les résultats que l'on pourrait considérer comme bons sont des cas où :

1. une seule proposition de traduction est faite et elle est correcte ;
2. plusieurs propositions sont faites et la première (après la classification en fonction de leur poids respectif) est la correcte ;
3. plusieurs propositions sont faites et la correcte n'est pas la première mais une autre dans la liste des résultats.

Pour la première méthode de prédiction, où l'on utilise les ensembles de traits qui décrivent les correspondances entre sens et équivalents, si l'on considère comme corrects les résultats des catégories (1) et (2), le rappel obtenu est de 59 % (ce qui signifie que des traductions correctes sont proposées pour 59 % des nouvelles occurrences) et la précision est de 83 % (83 % des propositions faites sont correctes). Si l'on considère comme corrects également les cas de la catégorie (3), on obtient un rappel de 66 % et une précision de 92 %.

Dans la deuxième méthode de prédiction, où l'on n'utilise que les graphes correspondant aux équivalents, les résultats sont meilleurs. Si l'on considère comme bons les résultats décrits par (1) et (2), le rappel obtenu est de 71 % et la précision est de 74 %. Si l'on inclut dans les résultats corrects les cas décrits par (3), on obtient un rappel de 91 % et une précision de 94 %.

Ces résultats divergents selon la méthode adoptée s'expliquent par les différents ensembles d'éléments auxquels le nouveau contexte est comparé. Dans la première méthode, le contexte des nouvelles occurrences est comparé aux ensembles d'éléments qui décrivent les correspondances sens-équivalents. Cet ensemble contient une sous-partie des éléments des graphes mis en relation, c'est-à-dire leurs éléments communs. Il peut arriver que les associations du contexte de la nouvelle occurrence ne soient pas trouvées dans cet ensemble. En revanche, avec la deuxième méthode, le contexte des nouvelles occurrences est comparé à toutes les associations dans les graphes des équivalents. Ces ensembles sont beaucoup plus grands que ceux utilisés dans la première méthode, alors, il n'y a que peu de nouvelles occurrences pour lesquelles une correspondance n'est pas trouvée.

On remarque que des propositions erronées sont faites surtout dans les cas d'équivalents très rares, où la quantité de segments correspondants dans le corpus d'entraînement sont très peu nombreux (de 1 jusqu'à 3). Ce problème est lié à la petite taille de notre corpus. Avec un corpus plus grand, où l'on aurait plus d'occurrences des équivalents rares, la performance de la méthode serait meilleure.

7 Discussion et perspectives

Les recouvrements et les divergences entre les contextes correspondant aux équivalents de traduction et aux sens des unités polysémiques repérés dans la LS soulèvent des questions concernant les relations qui peuvent exister entre les éléments de deux langues. La correspondance de plusieurs équivalents à un sens détecté dans la LS signifie-t-elle que l'on pourrait repérer des sous-sens au sein de ce sens, qui n'étaient pas mis en évidence pendant le processus de désambiguïsation des unités de la LS ? Ou plutôt, dans le cas où l'on parle d'usages dans la LS et pas de sens, la correspondance d'un équivalent à plusieurs usages pourrait-elle servir d'indice pour fusionner ces usages en un seul sens ?

Ce sujet doit être traité avec beaucoup de précautions. Il peut arriver que, dans la LC, il existe plus de distinctions sémantiques que dans la LS, surtout pour les mots ayant une portée sémantique vague au sein de celle-ci. Les distinctions mises en évidence par le calcul des

cooccurrents des unités de la LS et des équivalents peuvent ne pas toujours correspondre à des sens différents mais à des nuances de sens moins saillantes. Même si, parfois, on arrive à repérer des sens dans une langue par le biais d'une autre langue – sens qui n'étaient pas mis en évidence par le processus de désambiguïsation au sein de la première – la projection du découpage de l'espace sémantique d'une langue dans une autre ne conduit pas toujours à des distinctions pertinentes dans celle-ci. Cependant, ceci pourrait être d'une grande utilité pour les applications de TAL relatives à la traduction. Le repérage dans la LS des distributions contextuelles correspondant aux distinctions de sens propres à la LC et décrites par les équivalents différents peut faciliter la sélection des équivalents corrects pendant le processus de traduction. Dans ce cas-là, on est alors en quête des éléments discriminants du côté de la LS qui pourraient être liés aux variations de sens plus ou moins grandes dans la LC.

Le travail présenté dans cet article fait partie d'une étude plus vaste sur les questions de recouvrement et de divergences entre les éléments polysémiques de deux langues et des problèmes qui relèvent d'une entreprise de mise en correspondance de ceux-ci. Dans le travail à venir, nous envisageons d'incorporer une petite quantité d'informations linguistiques dans la méthode afin d'explorer les possibilités d'amélioration de nos résultats. Ces informations linguistiques peuvent concerner le repérage de termes complexes. On pourrait aussi répéter les expériences en éliminant les termes inhérents à la nature de notre corpus, c'est-à-dire les termes juridiques. L'élimination de ces termes laissera probablement apparaître des indices plus pertinents pour la distinction des sens différents des unités polysémiques ; les sens qui seront alors proposés correspondront bien plus à des sens qu'à des usages. Comme amélioration de la méthode nous projetons également la prise en compte des informations positionnelles des mots.

Une autre piste à explorer réside aussi dans l'étude de la similarité sémantique des équivalents différents d'un mot polysémique qui permettrait d'arriver à des conclusions concernant les distinctions de sens et le repérage de nuances de sens du côté de la LS. Les conclusions sur la similarité sémantique des équivalents pourraient aussi s'avérer utiles dans le processus de fusionnement des usages repérés en sens. Nous allons essayer de valider les sens et les sous-sens proposés en utilisant différentes mesures de similarité ou en ayant recours à des questionnaires auprès de locuteurs natifs des langues impliquées. En outre, une application de cette méthode à d'autres mots polysémiques nous permettra de dégager à l'avenir une évaluation plus globale.

Références

ALTENBERG B., GRANGER S. (2002). Recent trends in cross-linguistic lexical studies, dans ALTENBERG B., GRANGER S. (eds.), *Lexis in Contrast, Corpus-based approaches* (Amsterdam / Philadelphia: John Benjamins Publishing Company), 3-48.

APIDIANAKI M. (2005). Translation prediction using word cooccurrence graphs. Actes de *Corpus Linguistics 2005,* Birmingham, 14-17 juillet 2005 (http://www.corpus.bham.ac.uk/PCLC, à paraître).

BOURIGAULT D., AUSSENAC-GILLES N., CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA), " Techniques Informatiques et structuration de terminologiques*, dans PIERREL J.-M., SLODZIAN M., (eds.), Paris : Hermès. Vol. 18, n°1/2004, 87-110.

DAGAN I., ITAI A., SCHWALL U., (1991). Two languages are more informative than one. Actes de 29th *Annual Meeting of the Association for Computational Linguistics (ACL)*, Berkeley, California, 1991, 130-137.

DYVIK H. (2003). *Translations as a Semantic Knowledge Source,* Draft 2003. (http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/transknow.pdf).

FIRTH J.R. (1957). A Synopsis of Linguistic Theory, 1930-1955, dans *Studies in Linguistic Analysis,* Special Volume of the Philological Society (Oxford : Basil Blackwell), 1-32.

GALE W.A., CHURCH K.W., YAROWSKY D., (1993). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities* 26(5), 415-439.

GAVRILIDOU M., LABROPOULOU P., DESIPRI E., GIOULI V., ANTONOPOULOS V. and PIPERIDIS S., (2004) Building parallel corpora for eContent professionals. Actes de *MLR 2004, PostCOLING Workshop on Multilingual Linguistic Resources,* Geneva, 28 August 2004.

SALKIE R. (1997). Naturalness and contrastive linguistics, dans LEWANDOWSKA-TOMASZCZYK and P.J. MELIA (eds.) Actes de *PALC 97: Practical Applications in Language Corpora* (Lodz : Lodz University Press), 297-312.

SALKIE R. (2002). *Two types of translation equivalence*, dans ALTENBERG B., GRANGER S. (eds.) *Lexis in contrast: Corpus-based* (Amsterdam / Philadelphia : John Benjamins Publishing Company), 51-71.

SIMARD M. (2003). Translation Spotting for Translation Memories. Actes de *NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond,* Edmonton, Canada, 2003, 65-72.

TEUBERT W. (2002). The role of parallel corpora in translation and multilingual lexicography, dans ALTENBERG B., GRANGER S. (eds.), *Lexis in contrast: Corpus-based approaches* (Amsterdam / Philadelphia : John Benjamins Publishing Company), 189-214.

TRIANTAFYLLOU I., DEMIROS I., MALAVAZOS C., PIPERIDIS S. (2000). An alignment architecture for Translation Memory bootstrapping. Actes de *MT 2000*, Exeter, November 2000, 3.1-3.8.

TVERSKY A. (1977). Features of Similarity. *Psychological Review,* 84(4), 327-352.

VERONIS J., LANGLAIS P. (2000). Evaluation of parallel text alignment systems – The ARCADE project, dans VÉRONIS J. (ed.) *Parallel Text Processing* (Dordrecht: Kluwer Academic Publishers), 369-388.

VERONIS J. (2003). Hyperlex: cartographie lexicale pour la recherche d'informations. Actes de *TALN 2003*, Batz-sur-mer, 11-14 juin 2003, 265-274.

VIBERG Å. (2002). *Polysemy and disambiguation cues across languages: the case of Swedish* få *and English* get, dans ALTENBERG B., GRANGER S. (eds.), *Lexis in Contrast, Corpus-based approaches*, (Amsterdam / Philadelphia: John Benjamins Publishing Company), 191-150.

WIDDOWS D., DOROW B. (2002). A Graph Model for Unsupervised Lexical Acquisition, Actes de *19th International Conference on Computational Linguistics (COLING 19)*, Taipei, August 2002, 1093-1099.