Évaluation des taux de synonymie et de polysémie dans un texte

Claude de Loupy (1)

(1) Sinequa
loupy@sinequa.com
51-54, rue Ledru-Rollin
94200 Ivry-sur-Seine
France

Résumé – Abstract

La polysémie et la synonymie sont deux aspects fondamentaux de la langue. Nous présentons ici une évaluation de l'importance de ces deux phénomènes à l'aide de statistiques basées sur le lexique WordNet et sur le SemCor. Ainsi, on a un taux de polysémie théorique de 5 sens par mot dans le SemCor. Mais si on regarde les occurrences réelles, moins de 50 % des sens possibles sont utilisés. De même, s'il y a, en moyenne, 2,7 mots possibles pour désigner un concept qui apparaît dans le corpus, plus de la moitié d'entre eux ne sont jamais utilisés. Ces résultats relativisent l'utilité de telles ressources sémantiques pour le traitement de la langue.

Polysemy and synonymy are two basic problems for natural language processing. In this paper, an evaluation of the importance of these phenomena is presented. It is based on the semantic lexicon WordNet and its associated corpus SemCor. Thus, when there are, in average, 5 possible senses for each word in the corpus, only half of them are really used. Similarly, if 2,7 words can be used to designate a concept, more than half of them are never used. These results tend to put the usefulness of such a resource in perspective.

Keywords – Mots Clés

taux de polysémie, taux de synonymie, lexiques sémantiques, WordNet, SemCor
polysemy rate, synonymy rate, semantic lexicon, WordNet, SemCor

1 Introduction

La synonymie et la polysémie sont deux phénomènes fondamentaux des langues humaines qu'il convient de considérer avec attention dans presque toutes les applications de Traitement Automatique de la Langue (TAL). Ils sont essentiels au bon fonctionnement de la langue mais posent de grandes difficultés aux systèmes : en traduction automatique (traduction du mot anglais *bank* par *banque* ou *rive*), en recherche documentaire (sens de *table* dans « *table de logarithmes* » et « *table de cuisine* »), en synthèse de la parole (prononciation du mot *fils* dans la phrase « *Les fils de la couturière sont en soie* »). L'utilisation de lexiques sémantiques permet de résoudre (ou tenter de résoudre) certaines de ces difficultés. Mais quelle est

l'ampleur du problème ? Sommes-nous si souvent confrontés à la polysémie et à la synonymie ? Dans les sections qui suivent, nous présentons une évaluation statistique de ces phénomènes, à la fois par rapport à un lexique et par rapport à un corpus de texte. Pour cela, nous utilisons les informations présentes dans WordNet (Miller *et al.*, 1990) De plus, la signification des statistiques fournies est illustrées par leur impact dans le cadre d'une application type recherche documentaire (RD). En effet, l'utilisation de connaissances et de traitements sémantiques en RD est fortement controversée (CUSIRF, 2002).

2 La polysémie et WordNet

2.1 Présentation de WordNet

WordNet est un thesaurus pour l'anglais qui regroupe les termes par classes de synonymie et fournit des relations sémantiques (synonymie, hyponymie, méronymie, etc.). Les concepts définis dans WordNet, c'est-à-dire les groupes de synonymes ou synsets, sont au nombre de 91 591, correspondant à 126 525 lemmes. Le nombre de couples mots-sens (association entre un lemme et l'un de ses sens possibles) est de 168 141, c'est-à-dire qu'il y a, en moyenne, 1,3 sens par lemme dans WordNet. De plus, chaque terme est réparti au sein de 45 classes sémantiques (animaux, actions, etc.).
Bien sûr, WordNet présente bien des défauts et a été bien souvent critiqué. En particulier, dans le cas d'une RD, la finesse des sens (41 sens pour *run*) est très difficile à utiliser, bien que nous ayons montré certaines améliorations quand on utilise WordNet dans un tel système (Loupy, 2000 ; Loupy & El-Bèze, 2002). D'un autre côté, le mot *derby* est considéré non ambigu et renvoie au sens de *chapeau melon*. La course de chevaux est totalement ignorée (Schütze & Pedersen, 1995). Or, cette distinction est fondamentale si on considère un corpus de résultats sportifs sur les courses de chevaux. Par ailleurs, la représentation sémantique par les classes sémantiques est trop grossière. Néanmoins, WordNet présente déjà l'intérêt d'exister et permet d'aborder les textes selon un aspect sémantique qui n'était que difficilement accessible précédemment. Cela est d'autant plus vrai que la couverture de WordNet est relativement bonne (Loupy *et al.*, 1998). WordNet nous permet ainsi d'évaluer les taux de polysémie et de synonymie présents dans les textes. Les chiffres donnés par la suite ont été évalués à l'aide du SemCor, un corpus étiqueté manuellement avec WordNet.

2.2 Présentation du SemCor

Le SemCor (Miller *et al.*, 1993) est un extrait du Brown corpus de 171 documents différents contenant 197 360 mots dont 106 850 mots pleins, répartis en 11 182 phrases (la plus longue a 119 mots dont 65 mots pleins) et 3 056 paragraphes (le plus long contient 1 113 mots au total dont 537 mots pleins). Chaque mot plein a été étiqueté manuellement à l'aide des sens fournis dans WordNet. Selon le manuel fourni avec le corpus, le taux d'erreur de l'étiquetage serait de 13 % (SemCor, 1995). En fait, il est toujours difficile d'évaluer la qualité d'un étiquetage et nous précisons ces chiffres pour que le lecteur se fasse sa propre idée de la pertinence des résultats qui suivent. Nous pensons qu'ils restent valable malgré la confiance relative de l'annotation du fait du caractère très marqué des statistiques rapportées.

2.3 Importance de la polysémie

2.3.1 Évaluation du taux de polysémie dans le SemCor

En moyenne, il y a 1,3 sens par lemme dans WordNet. Cette ambiguïté n'est pas importante et l'on pourrait alors penser que la tâche de désambiguïsation est simple puisque l'ambiguïté est assez faible et que 82 % des entrées sont non ambiguës. Mais ce taux d'ambiguïté doit être évalué sur un corpus. Sur les 16 609 entrées WordNet utilisées dans le SemCor, seules 8 137 d'entre elles sont non ambiguës, et elles ne représentent que 23 171 occurrences (22 %) dans le corpus qui en comporte 106 850. Le rapport est donc inversé par rapport au lexique puisque l'on passe de 82 % de termes non ambigus à seulement 22 %. Le taux d'ambiguïté peut être estimé en rapportant la somme de tous les sens possibles au nombre d'occurrences des lemmes dans le corpus : $\frac{\sum_{\lambda} \mathrm{K}(\lambda) \cdot \mathrm{N}(\lambda)}{\sum_{\lambda} \mathrm{K}(\lambda)}$ où $K(\lambda)$ représente le nombre d'occurrences du lemme λ dans le SemCor et $N(\lambda)$ son nombre de sens dans WordNet. Ce qui donne une polysémie moyenne de 5 sens pour tous les lemmes et de 6,2 sens pour les lemmes ambigus.

2.3.2 Lien entre le nombre de sens et la fréquence d'occurrence

La Figure 1 indique la répartition des mots par nombre de sens dans WordNet et le SemCor.

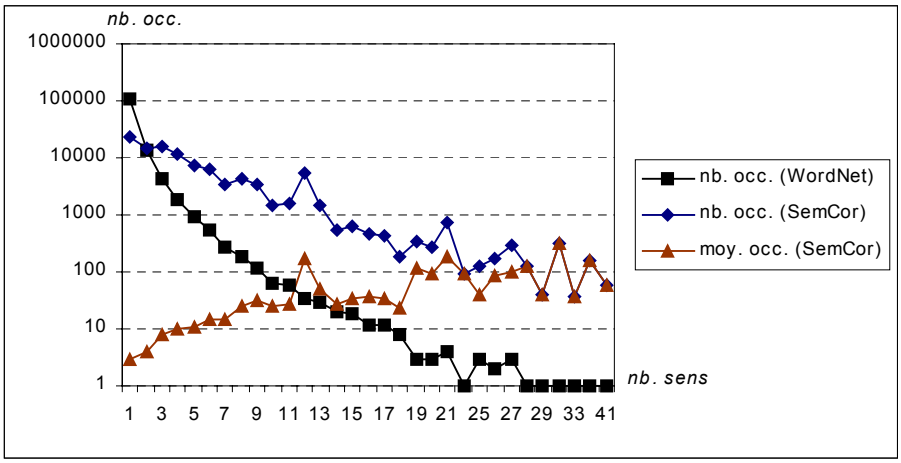

Figure 1 : Courbe de répartition des entrées WordNet selon leur nombre de sens

La première courbe *(nb. occ. - WordNet)* indique la répartition par nombre de sens dans WordNet. On voit qu'elle décroît très rapidement. La deuxième (*nb. occ. - SemCor*) indique cette répartition dans le SemCor (chaque occurrence d'un terme est comptée une fois). La aussi, la courbe décroît mais de façon beaucoup moins marquée. Enfin, la dernière courbe (*moy. occ. - SemCor*) indique le nombre moyen d'occurrences des termes en fonction de leur nombre de sens dans le SemCor. Cette fois, la courbe croît avec le nombre de sens (grossièrement). Cette courbe suit donc le principe de Zipf (1945) selon lequel plus un terme est fréquent, plus il a tendance à être ambigu. Or, la difficulté de la levée de l'ambiguïté sémantique dans le SemCor vient principalement de termes hautement polysémiques dont la fréquence est élevée : le verbe *run* a 41 sens différents et apparaît 59 fois et le verbe *have* a 21 sens et son nombre d'occurrences est 604.

Mais il convient de remarquer aussi que, en RD, on considère presque toujours que les termes les plus fréquents sont les moins informatifs, conformément aux théories de l'information classiques. Puisque les termes les plus fréquents sont les plus ambigus et que nous en tenons moins compte que les termes moins fréquents, il conviendrait de s'intéresser à des termes moins fréquents et donc moins polysémiques. Cela diminue donc la complexité de l'opération de désambiguïsation. Des expériences en ce sens seraient sans doute intéressantes.
Le Tableau 1 donne, dans le bloc de gauche, les 15 lemmes les plus fréquents (lemme, catégorie grammaticale, nombre de sens, fréquence d'occurrence dans le SemCor), dans le bloc central, les 15 couples lemme/sens les plus fréquents (lemme/sens, catégorie grammaticale, fréquence) et dans le dernier bloc, les 15 synsets les plus fréquents (synset, catégorie grammaticale, fréquence).

Lemme	**C**	**nb. sens**	**fréq.**	**Lemme, Sens**	**C**	**fréq.**	**Synset**	**C**	**fréq.**
be	v	12	4623	*person (être humain)*	n	3853	*person, individual, someone, mortal, human, soul*	n	3906
person	n	3	3853	*be (occuper une certaine position)*	v	2920	*be, occupy certain position, occupy certain area*	v	2920
not	r	1	928	*not (négation)*	r	928	*not, n't*	r	1306
group	n	3	838	*group (entités considérées comme une unité)*	n	838	*group, grouping*	n	838
location	n	4	613	*be (constituer, représenter)*	v	804	*constitute, represent, make up, be*	v	816
have	v	21	604	*location (un point ou une étendue)*	n	610	*location*	n	610
say	v	9	580	*say (exprimer une idée, etc.)*	v	504	*state, say, tell*	v	577
one	a	9	430	*have (posséder)*	v	437	*have, have got, hold*	v	445
n't	r	1	378	*one (une seule unité)*	a	419	*look, appear, seem*	v	298
man	n	10	356	*n't (négation)*	r	378	*man, adult male*	n	290
only	r	6	344	*two (cardinal)*	a	304	*one, 1, i, ane*		288
make	v	31	324	*man (une personne mâle adulte)*	n	290	*effect, carry out, make, do*	v	284
see	v	19	318	*be (être égal)*	v	269	*two, 2, ii*	a	282
more	r	3	318	*be (avoir lieu)*	v	269	*equal, be identical to, be*	v	272
two	a	1	304	*make (se lancer dans)*	v	263	*be, occur*	v	269

Tableau 1 : Lemmes, sens et synsets les plus fréquents dans le Semcor

Parmi les 15 lemmes les plus fréquents, la plupart figurent dans les 15 lemmes-sens et les 15 synsets les plus fréquents. Le verbe *be* (qui est le lemme le plus fréquent) apparaît dans 4 des 15 lemmes-sens les plus fréquents. De plus, il fait partie de 4 des 15 synsets les plus fréquents. En fait, si l'on considère le recouvrement entre les 15 lemmes et les 15 lemmes-sens les plus fréquents, on peut avoir une idée de l'effet de la polysémie dans les mots fréquents. Dans ce cas, on peut facilement calculer que, dans 83 % des cas, c'est le sens le plus courant qui est employé pour ces mots. De même, si l'on calcule le recouvrement entre les lemmes-sens et les synsets les plus fréquents, on peut avoir une idée de l'effet de la synonymie. On trouve alors que, pour les synsets les plus fréquents, le même représentant est utilisé dans 97 % des cas. Ce résultat tend à prouver que l'influence de la synonymie est faible par rapport à celle de la polysémie. Il convient de relativiser cette affirmation car elle n'est faite qu'à partir de l'observation des termes très fréquents.
Une conséquence de ces observations est que la répartition des sens et celle des mots sont intimement liées. On constate que les 4 premières occurrences de lemmes correspondent exactement aux 4 premières occurrences de lemmes-sens et aux 4 premières occurrences de synsets. Ces observations ont déjà été faites par Biber *et al.* (1994) et montrent qu'il est

important de connaître la répartition réelle des sens dans l'utilisation du mot en plus de connaître ses différents sens possibles.

2.3.3 Statistiques générales

Les tableaux 2 et 3 donnent certains résultats statistiques sur WordNet et le SemCor. Chaque information est évaluée sur tous les mots et par catégorie grammaticale.

	Tout	Adjectifs	Adverbes	Noms	Verbes
Nb. de lemmes	126 525	19 101	5 050	87 647	14 727
Nb.lemmes ambigus	22 283	5 303	771	11 515	4 694
Nb. de synsets	91 591	16 428	3 243	60 557	11 363
Nb. de mots-sens	168 217	28 762	6 203	107 484	25 768
Polysémie moy. (tous lemmes)	1,33	1,51	1,23	1,23	1,75
Polysémie moy. (lemmes ambigus)	2,87	2,82	2,49	2,72	3,35

Tableau 2 : Comportement statistique des synsets dans WordNet

	Tout	Adjectifs	Adverbes	Noms	Verbes
Nb. lemmes différents	17 098	4 043	1 308	8 451	3 296
Nb. lemmes ambigus	8 708	2 136	393	4 018	2 161
Nb. occurrences de lemmes	106 850	19 753	11 804	48 606	26 687
Nb. occurrences lemmes ambigus	82 688	14 195	6 565	37 513	24 415
Nb. de synsets	19 119	4 480	1 108	9 675	3 880
Nb. de mots-sens	23 930	5 317	1 627	11 379	5 607
Polysémie moy. (tous les lemmes)	5,02	4,25	2,56	3,98	8,58
Polysémie moy. (lemmes ambigus)	6,2	5,52	3,80	4,87	9,29

Tableau 3 : Comportement statistique des synsets dans le SemCor.

Les chiffres précédents montrent que le tableau (au sens figuré) que l'on peut dresser à partir d'un corpus est totalement différent de celui qu'on pourrait dresser à partir du lexique, en particulier en ce qui concerne la répartition des concepts et des ambiguïtés. Il y a 18,5 % d'adjectifs, 11 % d'adverbes, 45,5 % de noms et 25 % de verbes dans le SemCor (mots pleins uniquement). On peut aussi voir que la tâche la plus difficile est l'étiquetage des verbes puisque le taux d'ambiguïté dans cette catégorie est de 8,58 sens par verbes.

2.4 Taux réel de polysémie

Les statistiques données dans la section précédente permettent de savoir combien il y a de sens possibles par mot dans le SemCor. Mais il serait intéressant de savoir aussi combien de ces sens sont réellement utilisés. En effet, il ne suffit pas de savoir qu'un terme peut être très ambigu selon le lexique, il faut aussi se demander avec quelles fréquence les différents sens possibles sont utilsés. Le Tableau 4 donne les statistiques d'utilisation des sens des termes polysémiques présents dans le SemCor.
La deuxième colonne donne la proportion des sens d'un terme qui sont utilisés dans le SemCor. La troisième indique le pourcentage d'utilisation du sens le plus fréquent et la quatrième le minimum d'utilisation du sens le plus fréquent. On voit donc que moins d'un sens sur deux est réellement utilisé, ce qui rejoint les conclusions de Church et Mercer (1993) selon lesquelles la plupart des sens donnés dans un dictionnaire sont très peu utilisés. Ainsi, même si le nombre moyen de sens associés à un mot dans le SemCor plaide en faveur de l'utilisation d'un système de désambiguïsation sémantique, le taux réel de polysémie diminue l'espoir que l'on pourrait mettre dans une telle solution. En particulier, dans le cas d'une

recherche documentaire, une désambiguïsation sémantique ne serait utile que si le terme utilisé dans la requête est utilisé dans un sens qui n'est pas son sens le plus fréquent.

Catégorie grammaticale	Pourcentage de sens utilisés	Utilisation du sens le plus fréquent	Minimum d'utilisation du sens le plus fréquent
Tout	48,98 %	83,86 %	11,80 %
Adjs	45,79 %	86,58 %	23,08 %
Adv	65,99 %	84,17 %	36,84 %
Noms	49,75 %	84,91 %	20,69 %
Verbes	48,41 %	79,17 %	11,80 %

Tableau 4 : Utilisation réelle des sens dans WordNet.

Cela est particulièrement important si l'on considère les requêtes posées dans l'évaluation TREC (Harman, 1993). Les *titres* (1 à 5 mots sensés représenter le besoin informationnel) comporte des mots qui sont choisis pour être le plus informatif et le moins ambigu possible. Les conclusions mises en avant à l'aide de cet environnement montrent généralement le peu d'utilité d'outils de désambiguïsation sémantique et cela semble logique étant donnée la nature des requêtes. Malgré ce fait et malgré la non pertinence de WordNet pour une application en RD, il est tout de même possible d'obtenir une amélioration des performances en utilisant un système de désambiguïsation sémantique (Loupy & El-Bèze, 2002).

3 La synonymie et WordNet

3.1 Évaluation du taux de synonymie dans le SemCor

WordNet permet d'évaluer la difficulté que représente la synonymie. Le Tableau 5 indique le taux de synonymie (c'est-à-dire le nombre moyen de synonymes par mot) dans WordNet, évalué sur le SemCor lorsque le sens du terme est connu (ligne *Corp-desamb*) et lorsque ce sens n'est pas connu (ligne *Corp-amb*) en fonction de la catégorie grammaticale.

Catégorie gram.	Toutes	Adjectifs	Adverbes	Noms	Verbes
WordNet	1,84	1,75	1,91	1,77	2,27
Corp-désamb	2,68	2,14	3,32	2,51	3,08
Corp-amb	18,28	14,24	10,55	13,11	34,11

Tableau 5 : Taux de synonymie en utilisant WordNet

La connaissance du nombre de synonymes possibles lorsque le sens n'est pas connu est importante dans le cas de la recherche documentaire si l'on suppose un enrichissement sans désambiguïsation sémantique, c'est-à-dire sans connaître le sens des termes. La très grande différence entre les taux de synonymie lorsque l'on connaît le sens et lorsqu'il n'est pas connu se justifie tout à fait si l'on considère qu'un terme a, en moyenne, 5 sens dans le SemCor. Si l'on multiplie ce nombre par les 2,68 synonymes possibles par mot-sens, on se rapproche du taux de 18,28. On voit donc que le taux de synonymie moyen dans un texte est assez important puisque, selon ces chiffres, chaque mot pourrait être remplacé (selon WordNet) par presque deux autres si le sens est connu, par 18 autres sinon !
La Figure 2 montre la fréquence moyenne des lemmes dans le SemCor en fonction de leur nombre de sens : plus le terme est polysémique, plus il est fréquent en moyenne.

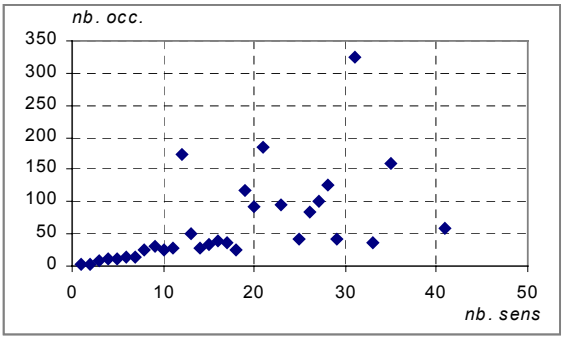

Figure 2 : Nombre moyen d'occurrences des termes dans le SemCor en fonction du nombre de sens.

La Figure 3 montre la fréquence moyenne des concepts (synsets) dans le SemCor en fonction de leur nombre de synonymes. S'il y a aussi une certaine tendance à l'accroissement de la fréquence en fonction du nombre d'éléments, en tout cas pour les groupes les moins nombreux, elle est beaucoup moins marquée. Nous reviendrons sur ce point avec le Tableau 6.
Enfin, la Figure 4 montre le nombre moyen d'occurrences des lemmes dans le SemCor en fonction de leur nombre de synonymes (tous sens confondus). La tendance de la courbe ainsi dessinée correspond à celle obtenue dans la Figure 2. On retrouve l'augmentation de la fréquence d'occurrence moyenne en fonction de l'augmentation du nombre de synonymes. On peut aussi constater qu'il y a une correspondance presque parfaite entre les deux courbes.

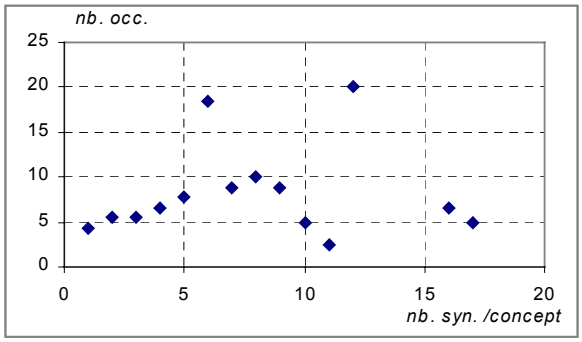

Figure 3 : Nombre moyen d'occurrences des mots-sens dans le SemCor en fonction du nombre de synonymes.

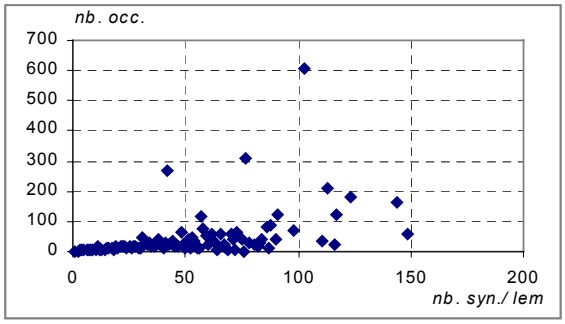

Figure 4 : Fréquence moyenne des termes dans le SemCor en fonction du nombre de synonymes (sens inconnu).

Cette correspondance est logique. Un terme appartient à autant de synsets qu'il a de sens. Chacun de ces synsets contenant un certain nombre d'éléments (en moyenne 2,68, cf. Tableau 5), le lien entre nombre de synonymes possibles et nombre de sens est immédiat. En revanche, il n'en va pas de même pour les synsets, ce qui est illustré dans le Tableau 6. Ce tableau donne les 15 lemmes qui ont le plus de sens (lemme, catégorie grammaticale, nombre de sens, fréquence dans le SemCor), les 15 lemmes ayant le plus de synonymes possibles sans tenir compte du sens (lemme, catégorie grammaticale, nombre de synonymes possibles, fréquence) et les 15 synsets ayant le plus d'éléments (synset, catégorie grammaticale, nombre d'éléments, fréquence). On pourra comparer ces statistiques à celles données dans le Tableau 1.
Sur ce tableau, on peut vérifier que les termes les plus ambigus sont aussi ceux qui ont le plus de synonymes possibles : entre les 15 lemmes les plus ambigus et les 15 lemmes ayant le plus

de synonymes, il n'y a que 4 termes de chaque côté qui n'apparaissent pas dans l'autre colonne. De plus, la fréquence d'occurrence est élevée puisque le lemme le moins fréquent dans le SemCor apparaît 11 fois (*check*) alors que la fréquence moyenne des lemmes est de 6,25. Pour les synsets, la fréquence moyenne est de 5,6 et si on considère les 8 concepts ayant 10 éléments ou plus, seuls 3 d'entre eux ont une fréquence supérieure à la moyenne. Pour ce qui est des concepts ayant 9 éléments, il y en a en fait 67. Seuls les 12 plus fréquents sont présents dans le Tableau 6. Parmi ces 67 concepts, seuls 21 apparaissent plus souvent que la moyenne. De plus, on peut aussi voir que les termes les plus ambigus et ceux qui ont le plus de synonymes sont très majoritairement des verbes. Cette tendance est un peu moins marquée pour les synsets. Enfin, il n'y a pas de correspondance entre les 20 lemmes les plus fréquents (cf. Tableau 1) et les 20 synsets ayant le plus d'éléments. Seuls *only* et *know* apparaissent dans l'un des synsets ayant le plus de synonymes. Cette constatation contredit l'affirmation de Voorhees (1993) selon laquelle les termes les plus fréquents sont ceux qui se trouvent dans les classes sémantiques les plus larges, c'est-à-dire celles ayant le plus d'éléments.

15 lemmes les plus ambigus				**15 lemmes ayant le plus de synonymes**				**15 synsets ayant le plus d'éléments**			
Lem.	**C**	**se**	**F**	**Lem.**	**C**	**syn**	**F**	**Synset**	**C**	**EL.**	**F**
run	v	41	59	run	v	149	59	botch, fumble, botch_up, muff, blow_it, flub, screw_up, ...	v	17	5
take	v	35	161	take	v	144	161	confuse, perplex, throw, fox, befuddle, fuddle, bedevil, ...	v	16	11
draw	v	33	36	pass	v	123	42	love, make_love, sleep_with, get_laid, have_sex, know, ...	v	16	2
make	v	31	324	make	v	123	324	three, 3, iii, trio, threesome, tierce, leash, troika, triad, ...	n	12	20
open	a	29	41	go	v	117	125	rebuke, rag, reproof, reprimand, jaw, dress_down, scold, ...	v	11	3
go	v	28	125	break	v	116	21	gorge, ingurgitate, overindulge, glut, englut, stuff, ...	v	11	2
give	v	27	210	give	v	113	210	bang-up, bully, cool, corking, cracking, dandy, great, ...	a	10	6
line	v	27	68	draw	v	111	36	dress_up, fig_out, fig_up, deck_up, gussy_up, fancy_up, ...	v	10	4
break	v	27	21	have	v	103	604	happen, hap, go_on, pass_off, occur, pass, ...	v	9	96
good	a	26	125	line	n	98	68	perfectly, plumb, completely, entirely, totally, utterly, ...	r	9	71
carry	v	26	41	good	a	91	125	presently, momentarily, in_a_moment, anon, soon, ...	r	9	46
play	v	25	63	open	a	90	41	find, happen_upon, chance_upon, hit_upon, bump_into, ...	v	9	43
pass	v	25	42	cut	v	88	12	dad, dada, daddy, old_man, pa, papa, pappa, pater, pop	n	9	39
lift	v	25	18	get	v	88	158	end, terminate, close_over, cease, run_out, stop, ...	v	9	28
head	v	23	94	check	v	87	11	traverse, track, cover, cross, pass_over, get_over, ...	v	9	19

Tableau 6 : Lemmes ayant le plus de sens, concepts et lemmes ayant le plus de synonymes dans le SemCor.

3.2 Taux réel de synonymie

Les statistiques données dans le Tableau 5 indiquent le nombre moyen de termes *possibles* pour désigner les concepts apparaissant dans le SemCor et le nombre moyen de synonymes qui existent pour tous les lemmes qui apparaissent dans le SemCor. Mais, en fait, ce qu'il conviendrait de savoir, c'est la propension à réellement utiliser des synonymes. En effet, un terme peut avoir un synonyme mais celui-ci peut ne jamais être utilisé. Le Tableau 7 donne des statistiques d'utilisation de synonymes dans le SemCor pour l'ensemble des lemmes et selon leur catégorie grammaticale. Ces données ont été calculées à partir des concepts utilisés dans le SemCor pour lesquels il existe plus d'un terme.

Catégorie grammaticale	Pourcentage de synonymes utilisés	Utilisation du synonyme le plus fréquent	Minimum d'utilisation du synonyme le plus fréquent
Tout	45,6 %	89,6 %	20,0 %
Adj.	42,9 %	89,4 %	20,0 %
Adv.	51,9 %	87,4 %	28,6 %
Noms	45,8 %	91,4 %	25,0 %
Verb.	45,9 %	87,4 %	25,0 %

Tableau 7 : Utilisation des synonymes dans le SemCor.

La deuxième colonne donne le pourcentage de termes utilisés pour désigner un concept rapporté aux nombre de termes possibles. On constate que 54,4 % des termes possibles ne sont pas utilisés. Cela est grandement confirmé par la troisième colonne. Elle donne le pourcentage d'occurrence du terme le plus fréquemment utilisé pour désigner un concept. Ainsi, on constate que, dans 89,6 % des cas, c'est le même terme qui est utilisé pour pointer un concept. Il s'agit d'un chiffre élevé et les possibilités de gain en terme de recherche documentaire sont donc limités. Il y a peu de possibilités d'accroissement des performances en RD par utilisation des synonymes si le terme utilisé dans une requête est le terme le plus fréquemment utilisé pour représenter un concept. Mais, bien sûr, il ne s'agit que d'une moyenne. Si on considère le concept *00014558*, il est représenté à égalité par ses deux éléments *shape* et *form* (5 fois chacun) dans le SemCor. Les performances sur une requête qui contiendrait *shape* pourraient être améliorées en utilisant aussi le nom *form*. Il convient donc de prendre en compte la répartition du concept sur ses différents éléments lors d'un enrichissement. La connaissance de la répartition des termes sur les concepts permettrait de déterminer quels enrichissements auront un impact important sur les performances. Cela est d'autant plus intéressant que nous savons, suite à de nombreuses publications (comme celle de Harman (1988)) que les performances sont améliorées lorsque l'utilisateur choisit un certain nombre de termes d'enrichissement parmi ceux que le système lui propose à partir d'un thesaurus. Il serait donc possible de limiter ces propositions aux seuls termes qui peuvent avoir un impact important, soit parce qu'ils apparaissent de nombreuses fois dans la collection, soit parce que le terme d'origine n'est pas majoritaire dans sa classe de synonymie.

4 Conclusion

Dans cette article, nous effectuons une évaluation des taux de polysémie et de synonymie présents dans un texte. Nous pouvons constater que, même si les taux de polysémie et de synonymie théoriques sont importants, le taux réel d'utilisation des différents sens d'un terme et des différents mots possibles pour exprimer un concept donné est relativement faible. De plus, ces deux phénomènes sont importants sur les mots ou les concepts les plus fréquents. Or, dans le cas de la recherche documentaire, par exemple, les termes les plus fréquents sont ceux qui apportent le moins d'information. Ils sont donc plus ou moins négligés.

Cette faible possibilité d'amélioration explique pourquoi les expériences d'utilisation de connaissances sémantiques en recherche documentaire donnent des résultats si contradictoires. On pourrait donc en conclure que ces deux phénomènes ne posent pas réellement de problème à un système de recherche documentaire (entre autre application). En fait, ce n'est pas le cas parce que, même si les cas où ces informations sont indispensables ne sont pas très fréquent, un utilisateur acceptera difficilement d'être confronté à des problèmes d'ambiguïté (bruit des moteurs) ou de synonymie (silence des moteurs) alors que ces questions lui semble évidentes. Et, même s'il l'accepte, c'est aux développeurs de tels systèmes de faire en sorte de lui faciliter la tâche. Il est donc nécessaire d'étudier chaque mot selon ses caractéristiques (nombre de sens, nombre de synonymes, fréquence, fréquence du

sens le plus fréquent, fréquence par rapport à celle du concept qu'il représente, etc.). En fonction de ces caractéristiques, des choix doivent être pris afin de choisir la meilleure stratégie à adopter. Un module de désambiguïsation sémantique peut alors être utilisé, ou un enrichissement par synonymie ou tout autre type de traitement sémantique selon le cas.
Ces conclusions doivent encore être confirmées par d'autres expériences, en particulier en utilisant d'autres lexiques. WordNet n'est peut-être pas le lexique le plus adéquat pour ce genre d'évaluations statistiques. Mais il faut aussi confirmer ces résultats sur d'autres corpus, d'autres langues, voire d'autres phénomènes sémantiques (hyponymie, etc.).

Références

Biber D., Conrad S., Reppen R. (1994), Corpus-Based Approaches to Issues in Applied Linguistics, in *Applied Linguistics*, vol 15, pp. 169-189.
Church K. W., Mercer R. L. (1993), Introduction to the special issue on computational linguistics using large corpora, *Computational Linguistics*, Vol. 19, No. 1, pp. 1-24.
CUSIRF (2002), *Creating and Using Semantics for Information Retrieval and Filtering State of the Art and Future Research*, Atelier en marge de *Third International Conference on Language Resources and Evaluation*.
Harman D. (1993), Overview of the First Text REtrieval Conference, *National Institute of Standards and Technology Special Publication 500-207*.
Harman D. (1988), Towards interactive query expansion, Actes de *11th Annual International ACMSIGIR Conference on Research & Development in Information Retrieval*, pp. 321-331.
Loupy C. de, El-Bèze M., Marteau P.-F. (1998), Word Sense Disambiguation using HMM Tagger, Actes de *First International Conference on Language Resources & Evaluation*, pp. 1255-1258.
Loupy C. de (2000), *Évaluation de l'apport de connaissances linguistiques en désambiguïsation sémantique et recherche documentaire*, Mémoire de Doctorat.
Loupy C. de, El-Bèze M. (2002, Managing Synonymy and Polysemy in a Document Retrieval System Using WordNet, Actes de *Creating and Using Semantics for Information Retrieval and Filtering State of the Art and Future Research*, Atelier en marge de *Third International Conference on Language Resources and Evaluation*.
Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K. (1990), Introduction to WordNet: An online lexical database, *International Journal of Lexicography*, Vol. 3 (4), pp. 235-244.
Miller G. A., Leacock C., Randee T., Bunker R. (1993) A semantic concordance, Actes de *3rd DARPA Workshop on Human Language Technology*, pp. 303-308.
Schütze H., Pedersen J. (1995), Information retrieval based on word senses, Actes de *4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-175.
Semcor man pages (1995), SemCor - Discussion of semantic concordance of semantically tagged text, http://www.cosgi.princeton.edu/~wn/man/semcor.7WN.html.
Strzalkowski T., Stein G. C., Wise G. B., Bagga A. (2000), Towards the Next Generation Information Retrieval, Actes de *RIAO'2000*.
Voorhees E. M. (1993), Using WordNet to disambiguate word senses for text retrieval, Actes de *16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 171-180.
Zipf G.K. (1945), Themeaning-frequency relationship of words, *Journal of general psychology 3*, pp. 251-256.