Extraction semi-automatique d'un vocabulaire savant de base pour l'indexation automatique

Lyne Da Sylva[1]

(1) École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Canada H3C 3J7
lyne.da.sylva@umontreal.ca

Résumé Le projet décrit vise à soutenir les efforts de constitution de ressources lexicales utiles à l'indexation automatique. Un type de vocabulaire utile à l'indexation est défini, le vocabulaire savant de base, qui peut s'articuler avec le vocabulaire spécialisé pour constituer des entrées d'index structurées. On présente les résultats d'une expérimentation d'extraction (semi-)automatique des mots du vocabulaire savant de base à partir d'un corpus ciblé, constitué de résumés d'articles scientifiques en français et en anglais. La tâche d'extraction a réussi à doubler une liste originale constituée manuellement pour le français. La comparaison est établie avec une expérimentation similaire effectuée pour l'anglais sur un corpus plus grand et contenant des résumés d'articles non seulement en sciences pures mais aussi en sciences humaines et sociales.

Abstract This project aims to help develop lexical resources useful for automatic indexing. A type of useful vocabulary for indexing is defined, the basic scholarly vocabulary, which can combine with specialized vocabulary items to form evocative, structured index entries. The article presents the results of an experiment of (semi-)automatic extraction of the basic scholarly vocabulary lexical items from a large corpus. The corpus is especially suited to the task; it consists of abstracts of scientific articles in French and English. The extraction task was successful in doubling the size of a previously manually compiled list. A comparison is made with a similar experiment conducted for English on a larger corpus which also contained summaries of articles in the humanities and social sciences.

Mots-clés : classes de vocabulaire ; indexation automatique ; extraction automatique ; corpus ; approche basée sur les corpus ; vocabulaire savant de base ; ressources lexicales ; français

Keywords: vocabulary classes; automatic indexing; automatic extraction; corpus; corpus-based approach; basic scholarly vocabulary; lexical resources; French

1 Introduction

Le présent article s'inscrit dans des travaux liés à l'indexation automatique. Il vise à soutenir les efforts de constitution de ressources lexicales utiles à l'indexation automatique. Dans notre approche à l'indexation, nous reconnaissons l'apport de deux types de vocabulaires utiles : le vocabulaire scientifique et technique spécialisé d'une part, et le vocabulaire « savant de base », qui s'articule avec le premier pour constituer des entrées d'index structurées complexes, plus utiles dans certains contextes d'indexation.

Nous présentons les résultats d'une expérimentation ayant pour but de créer une ressource lexicale, la liste du vocabulaire savant de base en français, constituée de mots répondant à des critères sémantiques spécifiques, esquissés ci-dessous. La création s'est appuyée sur l'extraction semi-automatique de données à partir d'un corpus ciblé. L'article présente la problématique, décrit l'expérimentation et ses résultats, et établit une comparaison avec une expérimentation similaire effectuée pour l'anglais.

2 Problématique

L'indexation traditionnelle peut s'envisager selon les deux approches suivantes : l'indexation macroscopique (*database indexing* en anglais, ou indexation pour les bases de données), où les documents sont décrits globalement par quelques termes d'indexation, permet de repérer un document dans une collection; l'indexation microscopique (*back-of-the-book indexing*, ou indexation de livres) sert à préciser pour un document donné les sujets abordés dans celui-ci et, surtout, dans quel passage les trouver. Dans chacune, la distinction ci-dessus entre le vocabulaire spécialisé et celui dit général se remarque, bien que de manière différente. Dans l'indexation macroscopique, des mots-clés thématiques comme « jurisprudence » ou « droit communautaire » (voir l'exemple (1) ci-dessous) sont à interpréter directement; alors que certains mots généraux, comme « application », doivent être interprétés en relation avec les mots thématiques (le document en question traite de l'application de la jurisprudence ou du droit communautaire, et non de la notion d'« application » en soi). Ces mots « généraux » seront définis comme le vocabulaire savant de base, ou VSB. Leur contribution doit ainsi être interprétée différemment.

(1) Droit communautaire, Application, Jurisprudence [1]

L'indexation microscopique formalise cette interprétation dans la structuration des entrées : une entrée d'index qui exprimerait la même notion que ci-dessus pourrait se présenter comme à l'exemple (2).

(2) droit communautaire, 186-189
application, 138, 141, 227

C'est dans le but de permettre la création d'entrées structurées que nous avons étudié les caractéristiques des vedettes principales et secondaires, et que nous avons dégagé cette distinction dans les types de mots qui y apparaissent.

L'utilisation différenciée des mots issus des différentes classes de vocabulaire est préconisée dans l'enseignement de l'indexation aux indexeurs professionnels. Le vocabulaire savant de base (VSB) est présenté chez Waller (1999). Elle le décrit comme étant situé entre le vocabulaire commun (acquis au cours de l'enfance) et le vocabulaire scientifique et technique spécialisé (relevant de la terminologie). Le VSB est acquis au cours des études secondaires et sert à exprimer le discours savant. On retrouve également une notion similaire dans les sections « mots-outils » des thésaurus documentaires, qui ne servent qu'à exprimer des aspects des termes du thésaurus. Les caractéristiques linguistiques du VSB peuvent être énumérées comme suit : ce sont des mots **abstraits** pour la plupart, ne désignant pas des objets concrets; ils appartiennent au vocabulaire **savant**, relevant d'un certain niveau d'érudition; enfin, ils sont très largement transdisciplinaires, avec un sens **général** et non spécifique à un domaine disciplinaire. Cette dernière caractéristique a inspiré notre expérimentation. Ces critères, bien que vagues et généraux, nous ont permis de décider si un mot doit être inclus dans le VSB ou non.

---

[1] Notice tirée de la base de données RESSAC, http://194.199.119.234/Ressac.htm, correspondant à l'article « Dans quelle situation, le droit de l'Union européenne trouve-t-il à s'appliquer en droit interne ? ».

Ce vocabulaire est difficile à circonscrire. On ne peut facilement se baser sur sa fréquence relative d'apparition dans les corpus : les termes « aspectuels » ne se distinguent pas a priori par leur fréquence d'apparition mais par leurs propriétés sémantiques. De plus, ils présentent la difficulté d'être polysémiques et ambigus : des mots comme « application » ont plusieurs définitions (notamment, le sens de « logiciel ») en plus de leur sens général (soit ici la nominalisation du verbe « appliquer », exprimant son processus ou son résultat). Leur fréquence d'utilisation peut donc être faussée par la coexistence des deux sens dans un document donné. L'objectif de la présente recherche est donc de trouver un moyen d'identifier un ensemble de mots qui relèvent de cette classe du VSB afin de guider un logiciel d'indexation automatique dans l'attribution d'entrées d'indexation.

3 Travaux antérieurs

Un certain nombre de chercheurs ont étudié la notion de classes de vocabulaire dans le but de caractériser un vocabulaire de base. Beaucoup d'entre eux relèvent du cadre de «l'anglais à des fins spécifiques » (Johns et Dudley-Evans, 1991). Certaines listes spécifiques ont été publiées. Ogden (1930) a développé le *Basic English* (anglais de base), pour les apprenants de l'anglais, constitué de 600 noms, 150 adjectifs, 100 mots outils et 18 verbes. La *Academic Word List* (AWL) (Coxhead, 2000) est conçue comme un vocabulaire spécifique de l'anglais pour fins académiques: Elle exclut les mots qui apparaissent dans les 2000 les plus fréquents de l'anglais, dont « capacity », « absence » ou « study » (que nous considérons du VSB). Certains chercheurs ont identifié les caractéristiques de l'anglais écrit scientifique et technique (ex. Swales, 1971). Celles-ci dépassent la notion de vocabulaire, abordant aussi des questions de style et de grammaire. Des travaux similaires ont été faits sur le français. Le « Vocabulaire général d'orientation scientifique » de Phal (1971) identifie un vocabulaire général axé sur la science, qui contient 1160 mots, dont des noms, adjectifs, verbes, etc., y compris des mots du vocabulaire commun. Drouin (2007) travaille à la définition d'un « lexique transdisciplinaire scientifique », utile pour les travaux de terminologie. Pour l'indexation, une liste du VSB ne doit contenir aucun mot du vocabulaire commun, ni spécialisé; il doit être limité aux noms (seuls utilisés en indexation), et les mots doivent couvrir toutes les disciplines, pas seulement les scientifiques (puisque l'indexation se fait autant dans les sciences sociales et humaines que dans les sciences pures et appliquées).

Très peu de travaux ont été consacrés à la nature des entrées d'index (c.f. Jones & Paynter, 2003; Wacholder & Song, 2003). Ces travaux n'ont cependant pas examiné les caractéristiques sémantiques ou lexicales des entrées d'index utiles. D'un point de vue théorique, Flaux et Van de Velde (2000) proposent une classification des noms, d'abord séparés en noms « véritables » (dénotant des choses) et noms « dérivés » (formés à partir de verbes ou d'adjectifs). Les mots du VSB appartiendraient vraisemblablement à la deuxième classe. Une étude plus approfondie de leurs critères de classement pourra apporter un éclairage sur les résultats obtenus par notre expérimentation.

Divers travaux portent sur l'extraction automatique de ressources lexicales, dont plusieurs en extraction terminologique (dont Vergne, 2005; Jacquemin et Bourigault, 2003; Daille, 1996). Cette optique est en fait à l'opposé de notre démarche, puisque nous cherchons des mots partagés par toutes les disciplines. Par ailleurs, les expressions à plusieurs mots renvoient davantage aux termes qu'au VSB, ce qui nous a orienté vers la recherche de mots simples (malgré l'existence d'exceptions comme « point de vue », multi-lexémique mais appartenant plutôt au VSB). Certaines approches opposent un corpus de référence (ou général) couvrant divers domaines à un corpus spécialisé, dont on veut extraire les termes (ou les expressions spécifiques) (voir par exemple Drouin, 2007).

4 Méthodologie expérimentale

Nous avons fait le pari qu'il était possible, au moins en partie, d'extraire automatiquement les éléments lexicaux qui nous intéressent à partir d'un grand corpus. Encore fallait-il le cibler de manière appropriée. Étant donné le type de mots visés, soit des mots du vocabulaire savant, d'application générale et de nature abstraite, nous avons cherché un corpus savant transdisciplinaire. Celui-ci devait être de taille suffisamment grande pour profiter des effets de nombre. Des travaux précédents effectués pour l'anglais (Da Sylva, 2010) nous avaient déjà menée sur la piste des bases de données bibliographiques.

Le corpus que nous avions constitué pour l'anglais comprenait des fiches bibliographiques contenant des titres et des résumés, de revues savantes variées: en beaux-arts, sciences sociales et humaines, sciences pures et appliquées. Le corpus était constitué d'environ 14 millions de mots anglais tirés de sept bases de données bibliographiques. Nous avons compté les fréquences d'occurrences des noms lemmatisés (à l'aide d'un dictionnaire et de règles morphologiques) dans ce corpus. Les noms dont la dispersion n'était pas suffisante parmi les sous-corpus ont été éliminés, ainsi que ceux qui sont exclusivement concrets dans WordNet. Les résultats obtenus sont très satisfaisants: sur les 1000 plus fréquents extraits, 664 mots ont été évalués comme appartenant effectivement à la classe du VSB selon nos critères sémantiques. Cette expérimentation nous a permis de quadrupler la liste initiale (166 mots) construite à partir de la caractérisation théorique du VSB anglais manuellement (par introspection et recherche de synonymes et analogies), de manière plus aisée. Pour le français, nous avons pu profiter d'un corpus similaire à celui que nous avions construit pour l'anglais, soit un ensemble de revues scientifiques publiées par l'ICIST[2]. Il s'agit de la même collection que celle décrite dans Nadeau et al. (2005). Elle couvre divers domaines scientifiques (voir le tableau 1). Bien que moins volumineux que celui pour l'anglais, il s'agit quand même d'un corpus savant transdisciplinaire, d'environ 2,5 millions de mots.

	Corpus français	% du corpus français	Corpus anglais	% du corpus anglais
Biochimie et biologie cellulaire	98 004	3.9%	86 668	4.0%
Revue canadienne de géotechnique	122 767	4.9%	105 624	4.9%
Revue canadienne de botanique	218 682	8.7%	196 278	9.1%
Revue canadienne de chimie	220 386	8.8%	192 165	9.0%
Revue canadienne de génie civil	121 500	4.8%	104 023	4.8%
Revue canadienne des sciences de la Terre	156 469	6.2%	132 521	6.2%
Journal canadien des sc. alieutiques et aquatiques	297 392	11.9%	255 171	11.9%
Revue canadienne de recherche forestière	345 294	13.8%	269 999	12.6%
Revue canadienne de microbiologie	173 754	6.9%	153 323	7.1%
Revue canadienne de physique	7 628	0.3%	10 470	0.5%
Revue canadienne de physiologie et pharmacologie	215 879	8.6%	183 357	8.5%
Revue canadienne de zoologie	343 125	13.7%	290 745	13.5%
Dossiers environnement	4 853	0.2%	6 421	0.3%
Géome	182 542	7.3%	160 705	7.5%
Total	2 508 275	100.00%	2 147 470	100.00%

Tableau 1 : Taille du corpus en nombre de mots

Pour identifier les noms concrets français, en l'absence d'une ressource comme WordNet (qui ne nous était pas disponible), nous avons combiné trois listes : la liste de mots concrets de Bonin et al. (2003), liste issue de travaux sur la « valeur d'imagerie » en psychologie et qui correspondrait selon les auteurs à

[2] Institut canadien de l'information scientifique et technique. Nous remercions Caroline Barrière, du Conseil national de recherche du Canada, d'avoir rendu possible le traitement de ce corpus.

la notion de concrétude (866 noms); une liste de noms d'animaux (*Les bestioles, insectes et animaux*, http://www.bestioles.ca/, 584 noms); et une liste de noms de métiers (pour identifier des humains) tirée de Wikipédia (*Liste des métiers*, http://fr.wikipedia.org/wiki/Liste_des_métiers, 781 noms). Ainsi, une liste de 2167 noms concrets a été créée (quelques doublons présents dans deux listes ont été éliminés).

Le corpus scientifique français fait 2 508 275 mots (environ 1/6 de celui utilisé dans l'expérience précédente – même ordre de grandeur bien que plus petit). Il est disponible en français et en anglais (traductions humaines), ce qui permet de comparer l'application de la méthode à notre corpus précédent. La méthodologie ci-dessus a été appliquée au corpus anglais: comptage des fréquences des noms lemmatisés du corpus (10 465 occurrences), en ne conservant que les noms présents dans au moins la moitié des revues et en éliminant les concrets (après le filtrage final: 1637 mots).

5. Résultats

Il est intéressant d'abord d'examiner les résultats par revue, puis ceux du corpus total, et enfin d'établir des comparaisons entre les différents corpus (en français, et pour les deux corpus en langue anglaise).

Revue canadienne de zoologie		Revue canadienne de recherche forestière		Revue canadienne des sciences alieutiques et aquatiques	
1496	femelle	1593	arbre	1449	poisson
1219	espèce	1475	croissance	1428	eau
1216	mâle	1360	peuplement	1022	lac
1039	rédaction	1201	forêt	984	rédaction
971	population	1047	espèce	970	modèle
916	taille	1020	rédaction	814	taux
755	cour	972	sol	788	population
583	étude	729	pin	711	croissance
567	taux	726	effet	707	taille
541	comportement	726	semis	678	effet
535	masse	724	modèle	678	espèce
529	résultat	682	site	670	donnée
528	nourriture	638	bois	643	pêche
524	effet	599	hauteur	628	concentration
511	habitat	579	coupe	594	cour
497	groupe	578	faible	537	saumon
484	petit	567	densité	506	être
476	nombre	566	taux	493	année
469	site	564	étude	475	stock
468	être	538	surface	473	âge
464	reproduction	537	être	463	étude
454	période	523	racine	454	abondance
445	différence	516	résultat	453	densité
422	zone	515	épinette	435	habitat
419	moyenne	506	feu	414	changement

Figure 1. 25 mots les plus fréquents de trois des sous-corpus scientifiques

Pour chaque revue, les mots les plus fréquents correspondent assez largement à des thématiques importantes du domaine (voir un échantillon à la figure 1). Bien sûr, certains de ces mots n'appartiennent pas au VSB. On trouve principalement ici trois types d'intrus : des mots spécialisés ou spécifiques (ex. « température », « population »), des concrets non identifiés comme tels à cause de la pauvreté de nos ressources (ex. « eau », « site ») et des mots ambigus avec une autre catégorie grammaticale (ex. « être »). Cette dernière classe a été traitée de manière spéciale, en l'absence d'un étiqueteur (*POS tagger*) : les mots à catégorie morphosyntaxique ambiguë ont été pour l'instant exclus de la liste extraite.

Si l'on considère maintenant l'ensemble du corpus, soit toutes les revues confondues, les mots les plus fréquents transcendent les thématiques disciplinaires et se généralisent, tel qu'espéré. Les 25 premiers

sont donnés à la figure 2, avec leur fréquence. Évidemment l'identification des concrets laisse à désirer : non seulement n'est-ce pas exhaustif (« eau » est un concret, comme l'indique WordNet pour « water »), mais aussi certains mots sont ambigus entre concret et abstrait; par exemple le nom d'animal « élan » est exclu sur la base qu'il apparaît dans la liste des concrets. Après analyse, sur les 1637 noms ainsi extraits, 682 mots sont jugés comme appartenant au VSB d'après les critères théoriques; cela représente près de 60% des mots non ambigus extraits (1150 mots sur 1637). Des statistiques: le VSB représente 100% des 6 premiers non ambigus; 90% des 100 premiers non ambigus; plus de 80% des 58 premiers (même avec ambigus); voir la figure 7.

4061	étude	2564	type	2048	région	1802	présence
3860	résultat	2505	faible	2046	structure	1757	méthode
3813	effet	2224	croissance	2013	concentration	1728	température
3689	espèce	2197	taux	1961	population	1676	augmentation
2865	analyse	2186	groupe	1869	valeur		
2817	eau	2184	modèle	1843	niveau		
2615	donnée	2060	site	1824	condition		

Figure 2. 25 mots les plus fréquents du corpus scientifique français total

4061	étude	1561	nombre	957	moyen
3860	résultat	1546	variation	948	longueur
3813	effet	1484	forme	935	milieu
3689	espèce	1417	différence	929	corrélation
2865	analyse	1365	changement	923	production
2615	donnée	1354	façon	893	masse
2564	type	1333	produit	858	expérience
2224	croissance	1313	activité	850	hypothèse
2197	taux	1291	moyenne	845	technique
2186	groupe	1260	relation	843	formation
2184	modèle	1225	zone	836	capacité
2048	région	1201	caractéristique	836	interaction
2046	structure	1199	développement	833	ensemble
2013	concentration	1195	mesure	825	distribution
1869	valeur	1185	base	811	potentiel
1843	niveau	1184	temps	809	traitement
1824	condition	1126	fait	807	mécanisme
1802	présence	1122	densité	798	utilisation
1757	méthode	1121	période	784	variable
1728	température	1073	réaction	775	fréquence
1676	augmentation	1029	cas	773	suite
1662	système	1026	partie	764	comportement
1656	surface	973	échantillon	757	résistance
1631	rapport	967	processus	754	état
1583	fonction	965	taille	746	rôle

Figure 3. 75 premiers mots du VSB extrait

La nouvelle liste de mots du VSB ainsi obtenue contient 682 mots, dont les 75 premiers sont présentés à la figure 3 (avec leur fréquence dans le corpus). Ces mots sont d'application générale, non spécialisée. En fait, cette liste doit être combinée à notre liste préalable, étant donné la faible couverture du corpus : seulement 1637 mots survivent à l'élimination des mots n'appartenant pas à au moins une moitié des revues; 64 mots originalement identifiés comme VSB n'étaient pas présents dans le corpus. La liste finale contient donc 746 mots. Notons que la liste originale (construite manuellement) faisait 430 mots.

Puisque le corpus que nous avons utilisé contient des résumés français et des résumés anglais pour les mêmes articles, nous avons comparé les résultats obtenus pour chacune des deux langues. Il sera ensuite intéressant de comparer les résultats obtenus ici pour l'anglais avec ceux que nous avions obtenus pour notre corpus anglais de 14 millions de mots, dans lequel davantage de disciplines sont représentées. La liste des 25 mots les plus fréquents du corpus scientifique anglais est donnée au tableau 2, première colonne. Parmi les mots n'appartenant pas au VSB, on note « species » et « population » (« found »,

« high » et « present » sont des intrus ambigus). En comparaison, les 25 mots les plus fréquents du grand corpus anglais y sont donnés en deuxième colonne. On voit que les mots n'apparaissent pas aux mêmes rangs; en outre, 11 mots seulement appartiennent aux deux listes (« study », « result », « effect », « analysis », « datum », « time », « structure », « group », « change », « system », et « type »).

Corpus scientifique anglais			Grand corpus anglais		
Fréq. ajustée	Mot	VSB?	Fréq. ajustée	Mot	VSB?
4703	study	1	17806	system	1
4362	result	1	17482	study	1
3817	species	0	15522	use	1
3776	effect	1	14569	analysis	1
3004	analysis	1	12894	reference	1
2754	level	1	11228	problem	1
2722	datum	1	11177	time	1
2659	rate	1	11081	process	1
2321	found	0	10987	form	1
2127	high	0	10682	datum	1
2082	concentration	1	10366	research	1
2079	population	0	10290	result	1
2068	time	1	9751	change	1
2059	structure	1	9697	relationship	1
2047	group	1	9690	effect	1
2032	change	1	9576	group	1
2008	system	1	9401	structure	1
1959	condition	1	9230	method	1
1956	increase	1	8702	present	0
1874	type	1	8658	source	1
1742	value	1	8483	information	1
1733	present	0	8004	theory	1
1719	temperature	1	7866	type	1
1664	area	1	7758	role	1
1648	region	1	7698	show	0

Tableau 2. 25 mots les plus fréquents du corpus scientifique anglais et du grand corpus anglais

Dans les premiers rangs de fréquence du corpus scientifique apparaissent plusieurs mots d'usage général dans un corpus scientifique et technique (plusieurs faisant référence à des calculs ou mesures : « rate », « concentration », « increase », « value ») qui ne sont pas également répandus dans des textes en sciences humaines et sociales ou en arts. Ceci renforce notre position que, pour identifier le VSB, il faut tenir compte de toutes les disciplines où l'on retrouve des publications savantes.

Regardons maintenant comment se comparent les corpus français et anglais équivalents. La figure 5 illustre que l'on trouve des pourcentages assez comparables au même rang dans l'analyse du corpus des deux langues (mais pour l'anglais, au rang 1176, 53% des mots sont du VSB, contre 63% pour le français). La figure 4 présente les 75 premiers mots (avec fréquence) extraits du corpus scientifique anglais et qui ont été jugés comme appartenant au VSB. On peut la comparer à la figure 3. Plusieurs équivalents traductionnels se retrouvent ou bien au même rang, ou bien à des rangs proches (notamment, les 3 premiers rangs reçoivent les mêmes mots). D'autres, dont la traduction est ambiguë, sont séparés (c'est le cas de « time », qui peut se traduire par « temps » ou « fois »). Le mot « structure » a presque la même fréquence d'occurrence dans les deux langues (2046 vs. 2059). Enfin, comparons les trois expériences. La figure 5 présente, par rang, le pourcentage des mots du VSB extraits du grand corpus anglais et des deux corpus scientifiques. L'analyse s'y était limitée aux 1000 premiers mots. Les

premiers rangs du grand corpus anglais étaient exempts de VSB, contrairement aux deux corpus scientifiques, ce qui en a fait un outil plus efficace de repérage de mots due VSB anglais. Les performances sur le corpus français sont meilleures que sur le scientifique anglais, sans doute dû à l'ambiguïté plus grande des mots anglais.

4703	study	1455	activity	855	production
4362	result	1407	factor	847	content
3776	effect	1391	size	840	characteristic
3004	analysis	1275	pattern	818	component
2754	level	1246	potential	801	evidence
2722	datum	1245	distribution	794	form
2659	rate	1234	range	775	parameter
2082	concentration	1225	variation	762	technique
2068	time	1185	process	759	behavior
2059	structure	1180	relationship	740	influence
2047	group	1119	treatment	735	estimate
2032	change	1110	mean	732	role
2008	system	1101	use	709	approach
1959	condition	1079	density	709	formation
1956	increase	1076	mass	691	frequency
1874	type	1056	length	688	decrease
1742	value	1053	presence	688	reduction
1719	temperature	1052	function	679	correlation
1664	area	1036	addition	675	variable
1648	region	1025	period	670	source
1645	response	1021	ratio	664	part
1633	method	1009	experiment	655	comparison
1633	number	957	field	652	average
1578	difference	911	interaction	633	stage
1456	control	872	sequence	610	contrast

Figure 4. 75 premiers mots du VSB extrait – corpus scientifique anglais

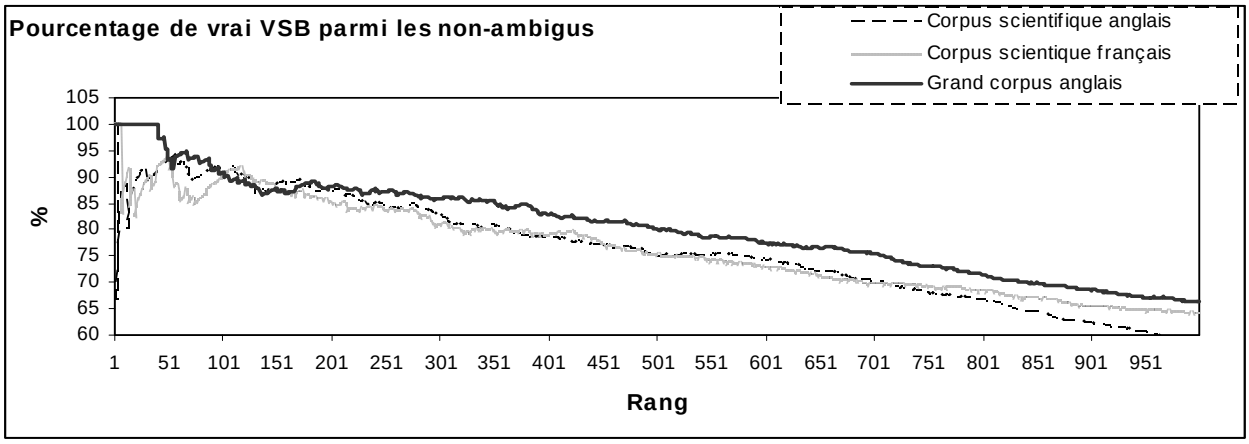

Figure 5. Pourcentage de vrai VSB par rang des noms extraits – comparaison des trois corpus

Avec le grand corpus anglais initial, 636 mots appartenant au VSB avaient été extraits (sur 1000 examinés). Avec le corpus scientifique anglais, 550 mots ont été extraits de la sorte (sur 1176 examinés); pour le corpus scientifique français, 682 (sur 1637 examinés). Sur ces deux derniers plus petits corpus, le VSB extrait est moindre en proportion. Le travail d'élimination de candidats non retenus a été plus grand, même pour l'anglais. En partie, les candidats examinés puis éliminés représentaient souvent des termes plus typiques des corpus scientifiques que des textes savants en général.

Il est intéressant de noter que dans tous les cas, on trouve parmi les hautes fréquences surtout des termes généraux; les termes thématiques ou disciplinaires sont beaucoup moins présents, ce qui est encore plus vrai dans le cas du grand corpus initial de l'anglais.

6. Discussion

On voit par l'analogie avec les résultats pour l'anglais que la liste obtenue pour le français est incomplète et biaisée par la nature scientifique et technique du corpus. Cependant, elle nous est quand même très utile, car elle constitue un excellent point de départ pour combler la liste manuelle établie jusqu'à présent. La liste extraite automatiquement contient en effet des mots absents de la première liste et en suggère d'autres, par analogie, dérivation, synonymes ou antonymes. Plus de la moitié des mots extraits appartiennent au VSB. Pour utiliser cette nouvelle liste, un travail de désambiguïsation sera nécessaire : les mots ainsi identifiés ont souvent plusieurs sens. Cela a d'ailleurs compliqué la tâche de déterminer si ces mots devaient ou non être conservés dans le VSB. Notre expérimentation pourrait bénéficier de la comparaison avec un corpus de langue générale (par exemple, journalistique). Nous sommes sceptique, cependant, quant à la distribution respective de mots tels que « application » et « début » dans chacun des corpus : elle risque de n'être pas très différente dans les deux. Mais l'hypothèse reste à vérifier. Il est possible par ailleurs que l'approche par comparaison de corpus puisse être utilisée « à l'inverse » : les mots qui sont exclus d'un corpus parce que n'appartenant pas au domaine pourraient être de bons candidats au statut de VSB. Diverses méthodes de calcul sont utilisées dans les recherches sur l'extraction de ressources lexicales. Pour le calcul de spécificité d'un mot par rapport à un corpus, une mesure comme le tf•idf peut être utilisée. Notre expérience n'a eu recours qu'à de simples fréquences et calculs de distribution, et pourrait être améliorée en considérant d'autres métriques.

Nous reconnaissons que la détermination de quels mots appartiennent au VSB constitue un travail subjectif, sans métrique objective de référence. Nous prétendons que cette classe de vocabulaire se définit en grande partie par son utilité dans l'indexation (à titre de vedette secondaire utile). Le réel travail d'évaluation de la justesse de ce VSB construit ne pourra se faire qu'en situation réelle, c'est-à-dire à l'intérieur d'un index. Da Sylva (2009) présente les situations d'utilisation des termes du VSB pour l'indexation. Il reste maintenant à construire un scénario d'évaluation qui permette de juger de la pertinence de distinguer les mots du VSB des mots du lexique technique et spécialisé d'un document. C'est la prochaine étape de notre programme de recherche.

7. Conclusion

Nous avons présenté les motivations derrière la création d'une ressource lexicale nouvelle ainsi qu'une expérimentation visant à la constitution semi-automatique de celle-ci. Force est de constater que l'apport de la présente expérimentation a été essentiellement d'accélérer le processus de constitution de la liste du vocabulaire savant de base, et non pas de l'automatiser entièrement. L'expérimentation a été un processus utile de découverte des éléments de vocabulaire utiles à l'indexation. Elle nous a permis également d'évaluer partiellement une liste obtenue pour l'anglais dans une expérimentation préalable.

Bibliographie

AUTEUR BONIN P., MÉOT A., AUBERT L., MALARDIER N., NIEDENTHAL P., CAPELLE-TOCZEK M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots. *L'année psychologique*, 103(103-4), 655-694.

COXHEAD A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2), 213-238.

DAILLE B. (1996). "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology". In Klavans J., Resnik P. (éds): *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 49–66. The MIT Press, Cambridge, Massachusetts.

DA SYLVA L. (2010). Corpus-based derivation of a "basic scientific vocabulary" for indexing purposes. In *Proceedings of the Corpus Linguistics Conference*, Univ. Liverpool, 21-23 juillet 2009.

DA SYLVA L. (2009). “Classes de vocabulaire et indexation automatique : le cas des index de livres”. In *Premier Workshop international sur la Terminologie et la sémantique lexicale (TLS’09)*, Université de Montréal, Montréal, 19-06-2009, 67-76. http://olst.ling.umontreal.ca/pdf/ProceedingsTLS09.pdf.

DROUIN P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 12(2), 45-64.

FLAUX N., VAN DE VELDE D. (2000). *Les noms en français : esquisse de classement*. Gap (France); Paris : Ophrys.

JACQUEMIN C., BOURIGAULT D. (2003). “Term Extraction and Automatic Indexing”. In *The Oxford Handbook of Computational Linguistics*, Oxford University Press.

JOHNS A.M., DUDLEY-EVANS T. (1991). English for Specific Purposes: International in Scope, Specific in Purpose. *TESOL Quarterly*, 25(2), 297-314.

JONES S., PAYNTER G.W. (2003). An Evaluation of Document Keyphrase Sets. *Journal of Digital Information*, 4(1). http://journals.tdl.org/jodi/article/view/93/92.

NADEAU D., BARRIÈRE C., FOSTER G. (2005) Bike: Bilingual Keyphrase Experiments. In *RANLP Workshop on Modern Approaches in Translation Technologies*, Borovets, Bulgaria.

OGDEN CK. (1930). *Basic English: A General Introduction with Rules and Grammar*. London: Paul Treber & Co., Ltd.

PHAL A. (1971). *Vocabulaire général d’orientation scientifique (V.G.O.S.) – Part du lexique commun dans l’expression scientifique*. Paris: Didier.

SWALES J. (1971). *Writing scientific English*. New York: Nelson.

VERGNE J. (2005). Une méthode indépendante des langues pour indexer les documents de l’internet par extraction de termes de structure contrôlée. In *Actes de la Conférence Internationale sur le Document Électronique* (CIDE 8), Beyrouth, Liban.

WACHOLDER N., SONG P. (2003). Toward a task-based gold standard for evaluation of np chunks and technical terms. In *Proceedings of HLT-NAACL 2003,* Edmonton, Canada, 189-196.

WALLER S. (1999). L’analyse documentaire : une approche méthodologique. Paris: ADBS.