MemLabor, un environnement de création, de gestion et de manipulation de corpus de textes

Vincent Perlerin

GREYC – CRNS UMR 6072 – Université de Caen
Campus II – BP 5186 – F14032 Caen Cedex
perlerin@info.unicaen.fr

Résumé – Abstract

Nous présentons dans cet article un logiciel d'étude permettant la création, la gestion et la manipulation de corpus de textes. Ce logiciel appelé *MemLabor* se veut un outil ouvert et *open-source* adaptable à toutes les opérations possibles que l'on peut effectuer sur ce type de matériau. Dans une première partie, nous présenterons les principes généraux de l'outil. Dans une seconde, nous en proposerons une utilisation dans le cadre d'une acquisition supervisée de classes sémantiques.

In this article, we present a study software that allows creation, management and handling of corpora. This software called *MemLabor* is an open-source program, adaptable to all operations that we can carry out on this type of material. In the first part of this article, we will present the main principles of this tool. In the second part, we will suggest one of it use within the framework of a semantic classes supervised acquisition.

Keywords – Mots Clés

analyse de corpus, acquisition supervisée de terminologie, sémantique lexicale
corpus analysis, supervised terminology acquisition, lexical semantics

1 Introduction

L'analyse automatique ou semi-automatique de corpus de textes a déjà montré son intérêt pour l'étude de phénomènes linguistiques ou l'acquisition terminologique. Des avancés notables ont eu lieux dans ce domaine tant du point de vue logiciel (Intex (Silbertein 1993), Hyperbase d'Etienne Brunet, Lexico d'André Salem…) que du point de vue des standards de représentation des corpus utilisés (le CES – Corpus Standard Encoding - crée à partir des recommandations de la TEI, les propositions du projet MATE pour la représentation des corpus de dialogues, le projet GENELEX…). Si ces standards sont d'une utilité certaine, leur complexité est souvent un frein à leur intégration aux logiciels d'étude pour l'analyse. Leur

vocation d'échange et de partage des résultats d'opérations effectuées sur les ensembles de documents en est considérablement affectée.

Dans cet article, nous présentons un environnement logiciel (*MemLabor*) permettant la création, la gestion et la manipulation de corpus de textes basés sur l'utilisation d'un format de conservation simple et léger (une DTD XML) pouvant intégrer à volonté d'autres propositions de standardisation par le biais d'URI[1] ou de chemins explicites sur des supports de stockage. Le logiciel assure la normalisation des documents : la possibilité est donc donnée d'intégrer à un corpus des documents directement issus de l'Internet (dans la version actuelle, aux formats HTML, XML et TXT). *MemLabor* se veut adaptable à tout type de travaux sur corpus de textes **(**annotations, recherche de patterns..) de par son caractère *open-source*[2] et sa vocation a conserver les documents dans leur état initial. Son utilisation dans le cadre de nos travaux vise la constitution supervisée de bases terminologiques structurées selon un modèle de sémantique différentielle et componentielle (Beust, 1998). Un des atouts majeurs du logiciel est de proposer un standard permettant de regrouper dans un même fichier XML, le corpus, les travaux effectués dessus et les relations entre ces entités. Ce standard est suffisamment ouvert pour permettre l'introduction de tout type de manipulations attendu que le chemin d'accès aux résultats et les caractéristiques de ces manipulations doivent être explicités au sein du fichier centralisant les documents du corpus.

2 Description du logiciel

Le logiciel proposé permet de gérer des corpus sous forme d'une description XML de fichiers de textes ou de fichiers de ressources linguistiques (mots grammaticaux d'une langue, lexies d'un domaine, étiquettes syntaxiques…) et des programmes qui prennent ces corpus en entrée et produisent des résultats (comptages, graphiques…) conservés avec le corpus. On peut y ajouter des programmes (classes, méthodes ou exécutables) tout en conservant les techniques de manipulation basées sur l'utilisation de la DTD XML proposée. Les fichiers peuvent être partagés entre plusieurs corpus puisqu'ils ne sont pas modifiés par les traitements et qu'ils peuvent être référencés par une URI ou tout autre chemin d'emplacement sur un support de stockage.

2.1 Création d'un corpus

Un corpus utilisable par le logiciel *MemLabor* doit être représenté par un fichier XML soumis à une DTD permettant un stockage d'information du type de la Figure 1. Cette DTD définit un document de type `CORPUS` déterminé par un nom, une suite de commentaires et une date de création (ligne 1). Chaque `CORPUS` est composé d'éléments de type `TRAVAIL_CORPUS` et `FICHIER_CORPUS`. Les éléments `TRAVAIL_CORPUS` correspondent à des travaux effectués sur l'ensemble des fichiers du corpus. Ils sont définis par un nom (le type de travail effectué sur le corpus), un fichier où sont stockés les résultats de ce travail (URI ou emplacement sur un support de stockage) et une date (celle à laquelle le travail a été effectué sur le corpus) (ligne

[1] Universal Resource Identifier – www.w3c.org/Addressing

[2] www.info.unicaen.fr/~perlerin

2). Les éléments de type `FICHIER_CORPUS` correspondent aux documents constituant le corpus : ils sont définis par un nom (URI ou emplacement sur un support de stockage) (ligne 3). Chaque document du corpus (de type `FICHIER_CORPUS`) peut être soumis à un travail particulier dont la trace sera sauvegardée dans le fichier XML du corpus selon le format défini pour les éléments de type `TRAVAIL_FICHIER` (ligne 4) où les attributs `Nom`, `Result` et `Date` correspondent respectivement au nom du travail correspondant, au fichier où sont stockés les résultats de ce travail et à la date de réalisation de ce travail.

```
1 <CORPUS Nom="Microsoft-Libé" Commentaires="Procès Microsoft 2001-2002"
Date="12/02/2002">
2 <TRAVAIL_CORPUS Nom="Zipf" Result="E:\Corpus\Zipf\Microsoft-Libé.zipf.xml"
Date="12/02/2002"/>
3 <FICHIER_CORPUS Nom="E:\Corpus\ Reprise des hostilités contre Microsoft.htm">
4 <TRAVAIL_FICHIER Nom="HTMLToTXT" Result="E:\Corpus\HTMLToTXT\ Reprise des
hostilités contre Microsoft.txt" Date="13/02/2002" />
5 </FICHIER_CORPUS>
6 </CORPUS>
```

Figure 1 : Exemple de document de type Corpus pour *MemLabor*.

La Figure 2 est une copie d'écran du logiciel *MemLabor*. Il s'agit de la première fenêtre d'interaction permettant de créer automatiquement un corpus (le fichier XML correspondant) en fonction d'un ensemble de fichiers choisis par l'utilisateur sur des supports de stockage. Le principe est identique pour des corpus constitués d'URLs externes, bien que l'utilisation d'URLs non personnelles puisse s'avérer problématique étant donnée la forte volatilité des documents du Web. Dans la partie gauche sont présentés les fichiers du répertoire sélectionné, dans celle de droite, les fichiers candidats au corpus en cours de création.

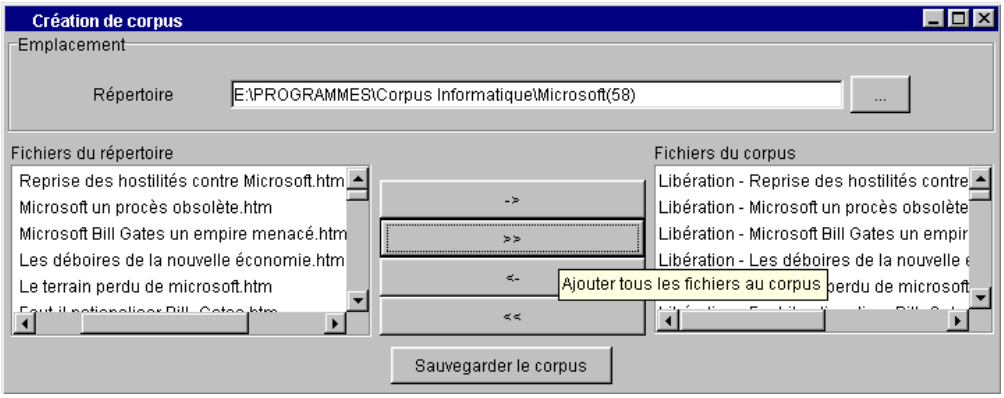


Figure 2 : Copie d'écran de *MemLabor* – création d'un corpus.

2.2 Travaux

Dans sa version actuelle, *MemLabor* propose cinq types de travaux sur corpus (ou sur des sous-ensembles du corpus) : la normalisation des documents d'un corpus au format TXT (depuis XML ou HTML), un segmenteur paramétrable (tokeniseur), un calcul de type Zipf (c.f. 2.2.1), une segmentation en paragraphe et une recherche de cooccurrences d'ensembles de lexies (c.f. 2.2.2).

2.2.1 Calcul de type Zipf

Zipf a observé (Zipf, 1949) que la fréquence d'utilisation des mots décroît de manière quasi-linéaire et que le produit f.R , soit la fréquence d'un mot multipliée par le rang de ce mot est à peu près constant (cette constante dépend du texte ou de l'ensemble de textes considéré). *MemLabor* permet d'effectuer un calcul de type Zipf sur l'ensemble des documents d'un corpus où sur un sous-ensemble de documents d'un corpus. Ce calcul donne lieu à la modification du fichier XML du corpus en fonction des travaux demandés et à la création d'un fichier de résultat rassemblant les lexies découvertes classées par ordre décroissant de leur nombre d'occurrences au sein des textes. Ce calcul est effectué à l'aide d'un segmenteur paramétrable permettant par exemple de prendre en compte ou non les mots composés contenant des tirets ou les groupes nominaux ou verbaux. Ce segmenteur utilise un moteur à base de règles déclaratives modifiable par les utilisateurs.

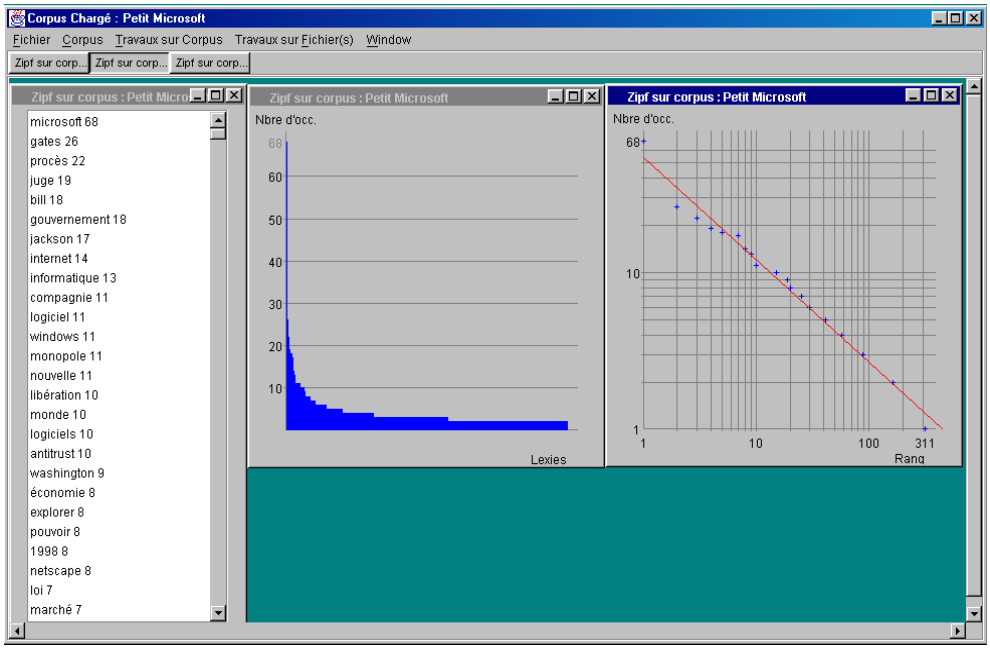


Figure 3 : Copie d'écran de *MemLabor* – calcul de type Zipf sur un corpus.

Lors d'un calcul de type Zipf, en plus de la liste des lexies (ou des motifs définies dans la base de règles du segmenteur) repérées avec leur nombre d'occurrences au sein de l'ensemble du corpus (fenêtre de gauche dans la Figure 3) sont proposées à l'utilisateur les représentations graphiques en histogrammes et en log/log des résultats (respectivement les fenêtres du centre et de droite dans la Figure 3). Ces graphiques permettent outre de vérifier la validité de la loi de Zipf sur le corpus considéré, de repérer d'éventuelles irrégularités inhérentes à des corpus hétérogènes (c'est-à-dire rassemblant des textes utilisant des vocabulaires très différents). Dans ce cas, la représentation en histogramme n'est pas d'allure logarithmique et la droite de régression – en rouge sur le graphique – de la représentation en log/log ne peut être significative étant donnée la dispersion des points du graphique.

2.2.2 Cooccurrences de lexies

L'utilisateur de *MemLabor* peut créer à l'aide du logiciel des fichiers de lexies (au format XML) correspondant à une DTD pré-définie fournie avec le logiciel. L'utilisation d'autres standards de bases terminologiques (TBX selon la norme ISO 12200[3], ISO DIS 16642 : TMF – Terminilogical Markup Framework, voir TC37/SC4) pouvant être prévue par l'utilisateur moyennant la programmation de modules adéquats.

Un ensemble de lexies (FIC_LEXIE) et leurs flexions peut être alors constitué selon le modèle suivant (Figure 4) :

```
1  <FIC_LEXIE Nom="Lexies en rapport avec Microsoft" >
2  <LEXIE Nom="Microsoft" Categorie="N">
3  <FLEXION>MS</FLEXION>
4  </LEXIE>
5  <LEXIE Nom="logiciel" Categorie="N">
6  <FLEXION>logiciels</FLEXION>
7  <FLEXION>software</FLEXION>
8  <FLEXION>softwares</FLEXION>
9  </LEXIE>
10 </FIC_LEXIE>
```

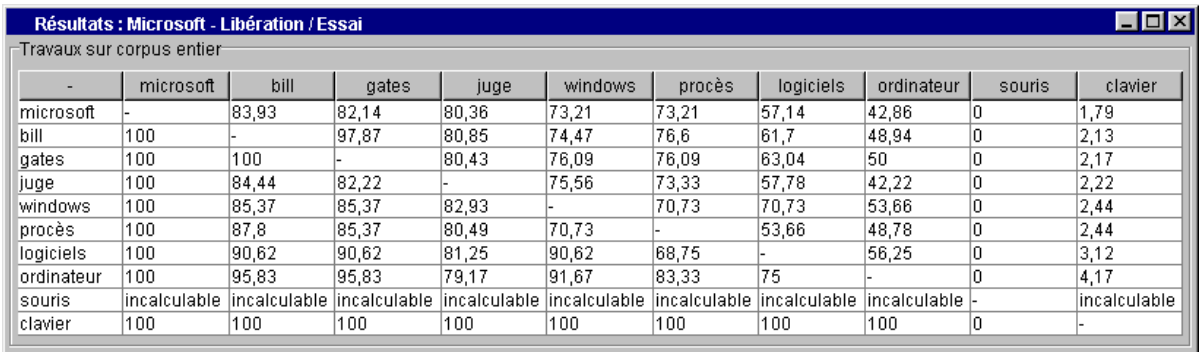
Figure 4 : Exemple de fichier d'ensemble de lexies.

Selon la DTD fournie avec le logiciel, une lexie doit être décrite par un nom et une catégorie. Les catégories proposées doivent être codées selon la norme Bdlex[4] (A pour les adverbes, C pour les conjonctions, N pour les noms…) car celle-ci peut ensuite être utilisée pour générer automatiquement les flexions de la lexie considérée à partir des données d'une base Bdlex. Les flexions seront codées dans le fichier XML selon le modèle de la ligne 6 de la Figure 4. L'utilisation de fichiers XML pour stocker ces données permet également, comme à la ligne 3 de la Figure 4, d'introduire des représentations textuelles sémantiquement identiques à la lexie (ici *MS* pour Microsoft) ne relevant pas des flexions courantes du mot ou des équivalents dans d'autres langues pouvant apparaître dans certains types de documents (lignes 7 et 8 de la Figure 4). L'utilisation étendue du terme « flexion » pourra alors être perçue comme erronée mais la représentation proposée pourra être utilisée pour des travaux sur corpus multilingues. Les fichiers de lexies ainsi constitués peuvent donner lieu à un calcul de cooccurrences au sein des documents d'un corpus choisi (Figure 5).

Préalablement aux calculs des cooccurrences des lexies, les documents du corpus peuvent faire l'objet d'une segmentation en paragraphes (repérage des indices typographiques dans les fichiers). Cette segmentation permet alors de limiter la prise en compte des cooccurrences à cette entité textuelle.

[3] www.lisa.org

[4] http://www.irit.fr/ACTIVITES/EQ_IHMPT/ress_ling/accueil01.php


Résultats : Microsoft - Libération / Essai

Travaux sur corpus entier

-	microsoft	bill	gates	juge	windows	procès	logiciels	ordinateur	souris	clavier
microsoft	-	83,93	82,14	80,36	73,21	73,21	57,14	42,86	0	1,79
bill	100	-	97,87	80,85	74,47	76,6	61,7	48,94	0	2,13
gates	100	100	-	80,43	76,09	76,09	63,04	50	0	2,17
juge	100	84,44	82,22	-	75,56	73,33	57,78	42,22	0	2,22
windows	100	85,37	85,37	82,93	-	70,73	70,73	53,66	0	2,44
procès	100	87,8	85,37	80,49	70,73	-	53,66	48,78	0	2,44
logiciels	100	90,62	90,62	81,25	90,62	68,75	-	56,25	0	3,12
ordinateur	100	95,83	95,83	79,17	91,67	83,33	75	-	0	4,17
souris	incalculable	incalculable	incalculable	incalculable	incalculable	incalculable	incalculable	incalculable	-	incalculable
clavier	100	100	100	100	100	100	100	100	0	-

Figure 5 : Copie d'écran de *MemLabor* – résultats du calcul de cooccurrences en pourcentage du nombre de fichiers du corpus pour un fichier contenant 10 lexies (ex : le mot *microsoft* apparaît dans 83,93% des fichiers où le mot *bill* apparaît dans l'ensemble des fichiers du corpus – le mot *souris* témoin dans l'expérience n'apparaît pas au sein du corpus).

3 Acquisition terminologique supervisée

3.1 Cadre de recherche - Problématique

Nos recherches en sémantique pour le TAL répondent aux exigences suivantes : nous désirons élaborer des modèles et construire des outils permettant des traitements syntagmatiques rapides et des représentations paradigmatiques non exhaustives *a priori*, c'est-à-dire aboutir à une sémantique *légère*. Nous nous plaçons dans une approche anthropocentrée où la machine se construit autour des besoins de l'utilisateur (Thlivitis, 1998**)**, et nous revendiquons une approche praxéologique de l'activité langagière **(**sémantique tournée vers la pratique de la langue par un individu ou un groupe restreint d'individus**)**. De plus, nous désirons constuire une sémantique lexicale intra-linguistique textuellement située (à l'opposé d'une sémantique purement référentielle qui étudierait les rapports des expressions au monde - selon la différence soulignée par (Auroux, 1999 p.38). Le logiciel décrit dans cet article et l'exemple d'application s'inscrivent donc pleinement dans ce cadre.

La réalisation de *MemLabor* est consécutive à une étude semi-manuelle débutée en 2000 dans le cadre d'un projet visant à proposer des outils de filtrage et de réordonnancement de résultats de systèmes documentaires classiques (moteurs de recherche du Web) en fonction de ressources sémantiques fournies par l'utilisateur (Perlerin, 2001). Cette étude semi-manuelle de 1783 dépêches journalistiques (Corpus *Reuter*) avait montré que les pourcentages de cooccurrences en nombre de fichiers à l'intérieur d'un corpus homogène (Figure 5) pouvaient être un indice valable pour le choix de lexies appartenant à des thèmes proches en vue de construire des systèmes de classes de catégorisation selon le modèle Anadia (Coursil, 2000)(Beust, 1998). Cette observation peut être partiellement expliquée par la notion d'isotopie (récurrence syntagmatique d'un même trait sémantique) support crucial du sens dans les textes. Les lexies candidates à une même classe de catégorisation partagent en effet un certain nombre de traits génériques (Rastier, 1994) dont la redondance au sein des entités textuelles conditionne le sens. C'est la détermination du global sur le local.

3.2 Aide pour la constitution de classes de catégorisation sémantique

Nos recherches concernent la dimension thématique de la cohésion textuelle. A partir d'un corpus homogène, il s'agit ici de constituer des classes de catégorisation sémantiques au sens des *taxèmes* de la sémantique interprétative (Rastier, 1994) : « *structure paradigmatique constituée par des unités lexicales se partageant une zone commune de signification et se trouvant en opposition immédiate les unes avec les autres* ». En d'autres termes et dans un cadre différentiel, il s'agit de structurer des ensembles de mots appartenant à un même thème en formant des sous-ensembles au sein desquels on peut marquer leur différence - voir le modèle Anadia (Nicolle et al., 2002). Notre but est de fournir aux utilisateurs une aide à la constitution de telles ressources.

Deux tâches sont à réaliser : l'extraction des candidats termes et le classement de ces candidats aux seins de sous-ensembles. Dans la littérature, on trouve de nombreuses techniques d'acquisition de terminologie : en fonction de critères principalement morpho-syntaxique (TERMINO [David et Plante, 1990], et LEXTER [Bourigault, 1994], …), en fonction de critères principalement statistiques (ANA [Enguehard, 1993], [Riloff et Shreperd, 1997],…), ou encore en fonction de marqueurs *a priori* (SEEK de Christophe Jouis, COATIS de Daniela Garcia). Désirant restreindre au maximum la quantité de ressources nécessaires au traitement et placer l'utilisateur au cœur du système, nous avons opté pour l'extractions des mots du domaine pour une technique statistique simple. Nous nous basons sur un calcul de type Zipf filtré à l'aide d'une liste de mots sémantiquement limités. En effet, lors du repérage des lexies mises en jeu dans les documents d'un corpus homogène, les graphiques en histogrammes obtenus présentent trois groupes consécutifs (numérotés 1, 2 et 3 dans la Figure 6). Le premier rassemble les lexies fortement redondantes dans la langue et n'ayant pas un potentiel sémantique important (déterminants, pronoms, articles…). Le second regroupe principalement les mots spécifiques au domaine du corpus. Le troisième ne présente pas de particularités remarquables.

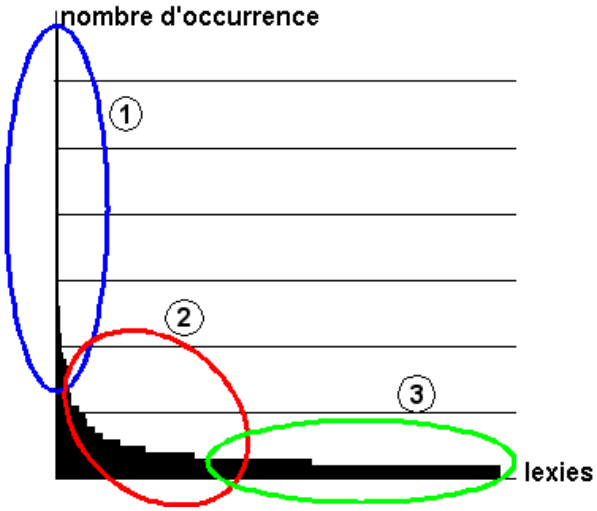
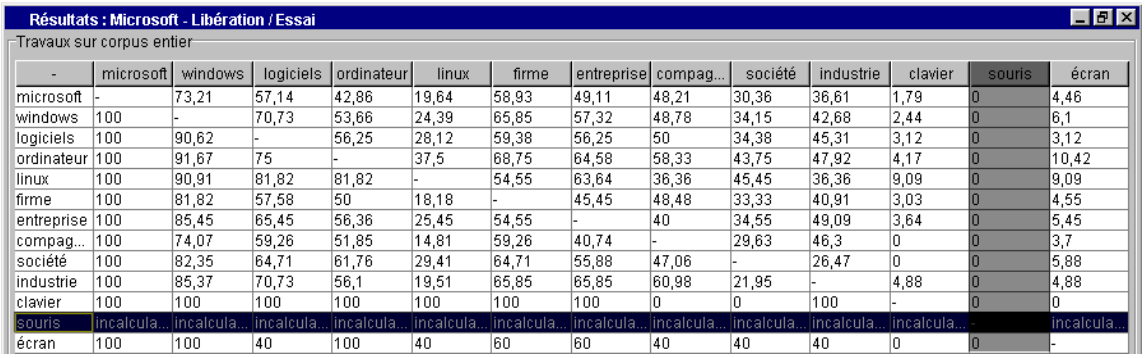


Figure 6 : Représentation graphique d'un calcul de type Zipf sur un corpus homogène.

MemLabor est actuellement disponible avec une liste de mots ayant un potentiel sémantique faible ou *stoplist* en français contenant 744 entrées. Cette liste représente 6Ko de mémoire et contient essentiellement les déterminants, pronoms, chiffres, auxiliaires et adverbes courants.

Des *stoplists* d'autres langues sont disponibles sur le web (ex : http://download-west.oracle.com/otndoc/oracle9i/901_doc/text.901/a90121/astopsup.htm#1234).

En ne tenant pas compte des mots présents dans la *stoplist* lors d'un calcul de type Zipf, la liste des lexies obtenue rassemble majoritairement des mots spécifiques au domaine du corpus (fenêtre de gauche dans la Figure 3). Parmi cette liste, l'utilisateur peut regrouper les mots qui selon sa pratique de la langue relèvent de la même thématique. Pour l'aider dans la constitution des classes de catégorisation, le calcul des cooccurrences des lexies sélectionnées pour un thème avec les autres lexies de la liste (comme dans la Figure 5) permet au logiciel de lui proposer la liste des lexies fortement cooccurrentes avec celles déjà présentes dans la classe qu'il est en train de constituer. Sur un corpus homogène et pour un domaine thématique donné, la prise en compte du document dans son entier comme surface d'exploration des cooccurrences peut s'avérer suffisante pour aider l'utilisateur dans son choix des ensembles de catégorisation.



Résultats : Microsoft - Libération / Essai

Travaux sur corpus entier

-	microsoft	windows	logiciels	ordinateur	linux	firme	entreprise	compag...	société	industrie	clavier	souris	écran
microsoft	-	73,21	57,14	42,86	19,64	58,93	49,11	48,21	30,36	36,61	1,79	0	4,46
windows	100	-	70,73	53,66	24,39	65,85	57,32	48,78	34,15	42,68	2,44	0	6,1
logiciels	100	90,62	-	56,25	28,12	59,38	56,25	50	34,38	45,31	3,12	0	3,12
ordinateur	100	91,67	75	-	37,5	68,75	64,58	58,33	43,75	47,92	4,17	0	10,42
linux	100	90,91	81,82	81,82	-	54,55	63,64	36,36	45,45	36,36	9,09	0	9,09
firme	100	81,82	57,58	50	18,18	-	45,45	48,48	33,33	40,91	3,03	0	4,55
entreprise	100	85,45	65,45	56,36	25,45	54,55	-	40	34,55	49,09	3,64	0	5,45
compag...	100	74,07	59,26	51,85	14,81	59,26	40,74	-	29,63	46,3	0	0	3,7
société	100	82,35	64,71	61,76	29,41	64,71	55,88	47,06	-	26,47	0	0	5,88
industrie	100	85,37	70,73	56,1	19,51	65,85	65,85	60,98	21,95	-	4,88	0	4,88
clavier	100	100	100	100	100	100	100	0	0	100	-	0	0
souris	incalcula...	incalcula...	incalcula...	incalcula...	incalcula...	incalcula...	incalcula...	incalcula...	incalcula...	incalcula...	incalcula...	-	incalcula...
écran	100	100	40	100	40	60	60	40	40	40	0	0	-

Figure 7 : Copie d'écran de *MemLabor* - résultats partiels du calcul des cooccurrences de 35 lexies en pourcentage du nombre de fichiers d'un corpus de 150 articles du journal Libération sur le procès Microsoft en 2001 (la lexie *souris* est absente du corpus).

Soit L_i, la lexie i et $P(L_i,L_j)$ le pourcentage en nombre de fichiers au sein du corpus contenant L_i où L_j apparaît[5]. Si les lexies L_1, L_2, L_3 …ont été sélectionnées comme candidates à une même classe sémantique, le logiciel proposera en fonction d'une table de résultats identique à celle présentée dans la Figure 7, une liste de lexies L_i classées en fonction du produit $P(L_1, L_i)xP(L_2,L_i)_{...}$ le facteur multiplicateur permettant de rendre compte des proportions de cooccurrence.

Une première expérience (décrite dans (Nicolle et al., 2002)) sur le corpus *Reuter* a montré que les résultats ainsi obtenus plaçaient les mots initialement choisis pour être candidats à une même classe sémantique parmi approximativement les 15 premières places de la liste. Dans le cas du corpus utilisé pour cet article (corpus Libération – cf. Figure 7), la Figure 8 présente la liste obtenue pour les lexies déjà candidates à une même classe : *entreprise* et *firme* (le tableau de cooccurrences ayant été calculé pour les 60 premières lexies de la liste obtenue suite au calcul de type Zipf[6]).

---

[5] dans la Figure 7, P(microsoft,linux) = 19,64 % et P(linux,microsoft) = 100%.

[6] l'ensemble des données de l'étude sont disponibles sur www.info.unicaen.fr/~perlerin

	firme	*entreprise*	**Produit**
logiciel	68	74	**5032**
internet	100	50	**5000**
industrie	**65,85**	**75,61**	**4978,9185**
gouverment	73,53	61,76	**4541,2128**
justice	71,43	63,49	**4535,0907**
société	**64,71**	**67,65**	**4377,6315**
produit	70,59	61,76	**4359,6384**
actionnaire	**61,11**	**69,44**	**4243,4784**
informatique	66,67	63,33	**4222,2111**
logiciel	59,38	70,31	**4175,0078**
concurrent	66,67	61,9	**4126,873**

	firme	*entreprise*	**Produit**
gates	60,87	65,22	**3969,9414**
verdict	57,89	68,42	**3960,8338**
bill	61,7	63,83	**3938,311**
juge	68,18	56,82	**3873,9876**
appel	71,43	53,57	**3826,5051**
microsoft	60	61,82	**3709,2**
procès	65,85	56,1	**3694,185**
pc	57,14	64,29	**3673,5306**
antitrust	**60,71**	**58,93**	**3577,6403**
monopole	**58,06**	**56,45**	**3277,487**
communication	59,26	53,7	**3182,262**

Figure 8 : Extrait de la liste des lexies proposées pour les lexies *firme* et *entreprise*. (La liste complète contient 60 lexies).

Comme on peut le constater dans la Figure 8, les mots raisonnablement candidats à une même classe sémantique que *firme* et *entreprise* (en gras dans la liste), ne sont pas classés aux premières places mais apparaissent à des positions cohérentes, plus rapidemment atteignables par l'utilisateur que dans la liste Zipf globale. L'expérience a été menée sur un corpus non segmenté[7] et ne saurait être validée que dans des conditions réelles. Une expérience sera ménée lors de l'atelier-formation organisé par l'ARCO[8] et le CNRS en juillet de cette année[9]. D'une manière générale, nous pensons que ce type de calculs supervisés par un utilisateur à même d'investir une partie de son temps à la constitution de telles ressources pourrait, couplés avec d'autres traitements légers (comme la reconnaissance des noms au sein des documents), représente une aide importante. Notre objectif est de proposer des outils sémantiques ne nécessitant pas de ressources paradigmatiques importantes ; ni de traitements préalables importants du corpus. L'analyse distributionnelle de documents basée sur une observation de caractéristiques linguistiques (redondances des termes, principe d'isotopie, …) est une piste intéressante de recherche pour aider à la constition de telles ressources.

4 Conclusion

Dans cette article, nous avons proposé une plateforme de gestion de corpus pour le TAL. Le logiciel *MemLabor* s'inscrit dans une approche coopérative de la recherche en TAL en permettant l'échange de corpus ou de résultats obtenus sur ces corpus par l'intermédiaire d'une DTD XML. Le caractère *open-source* du programme en assure aussi sa possible d'évolution.

Cette ressource logicielle exploite et rend compte de la dimension intertextuelle des documents. Des études linguistiques récentes (Rastier, 2001) montrent l'importance dans le processus d'interprétation de la place d'un document dans un ensemble saisi par le lecteur

[7] des résultats sur des documents segmentés (pourcentages de cooccurrences dans les segments de documents) sont disponibles sur www.info.unicaen.fr/~perlerin

[8] Association pour la recherche cognitive

[9] http://users.info.unicaen.fr/~anne/HTML/atelier.htm

(ici, le corpus constitué par l'utilisateur) . *MemLabor* a pour objectif de participer à ce genre d'étude, (re)plaçant le document au sein d'entités plus grandes influençant son interprétation et permettant des expérimentations sur cette dimension textuelle.

Références

Auroux S. (1999), *Le langage, la raison et les normes*, Paris, PUF.

Beust P. (1998) *Contribution à un modèle interactionniste du sens*, Thèse de Doctorat en Informatique de l'Université de Caen.

Bourigault D. (1994), *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*. Thèse en informatique linguistique, Ecole des hautes Etudes en Sciences Sociales, Paris.

Coursil J. (2000), *La fonction muette du langage - Essai de linguistique générale contemporaine*, Editions Ibis Rouge.

David S, Plante P. (1990). *De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes*, ICO, 2(3):140-154.

Enguehard, C (1993). *Acquisition de terminologie à partir de gros corpus*, Informatique & Langue Naturelle, ILN'93, Nantes,. 373-384.

Nicolle A, Beust P, Perlerin V. (2002), Un analogue de la mémoire pour un agent logiciel interactif, *In Cognito* N°21, 37-66.

Perlerin V (2001), La recherche documentaire : une activité langagière, Actes de *TALN-RECITAL 2001*, 469-479.

Rastier F. (2001), *Eléments de théorie des genres*, texte diffusé sur la liste fermée *Sémantique des textes*, 2001. http://www.atala.org/je/010428/Rastier/Rastier280401.html

Riloff E., Shepherd J. (1997). A corpus-based approach for building semantic lexicons. In Cardie, C. et Weischedel, R., editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 117-124. ACL, Somerset, New Jersey.

Silberstein M. (1993), *Le système INTEX, Dictionnaires électroniques et analyse automatique de textes*, Paris, Masson.

Thlivitis T. (1998), *Sémantique interprétative Intertextuelle : assitance informatique anthropocentrée à la compréhension des textes*, Thèse de l'Université de Rennes 1.

Zipf G. K. (1949), *Human Behavior and the Principle of Least Effort*, Addison-Wesley.