Acquisition de relations lexicales désambiguïsées à partir du Web

Chrystel Millon

Equipe DELIC – Université de Provence
29, Av. Robert Schuman – 13621 Aix-en-Provence Cedex 1
Krystelkay@aol.com

Résumé – Abstract

Nous montrons dans cet article qu'un pré-étiquetage des usages des mots par un algorithme de désambiguïsation tel qu'*HyperLex* (Véronis, 2003, 2004) permet d'obtenir des relations lexicales (du type *NOM-ADJECTIF, NOM de NOM, NOM-VERBE*) beaucoup plus exploitables, parce qu'elles-mêmes catégorisées en fonction des usages. De plus, cette technique permet d'obtenir des relations pour des usages très peu fréquents, alors qu'une extraction indifférenciée « noie » ces relations au milieu de celles correspondant aux usages les plus fréquents. Nous avons conduit une évaluation sur un corpus de plusieurs milliers de pages Web comportant l'un des 10 mots-cibles très polysémiques choisis pour cette expérience, et nous montrons que la précision obtenue est très bonne, avec un rappel honorable, suffisant en tout cas pour de nombreuses applications. L'analyse des erreurs ouvre des perspectives d'améliorations pour la suite de notre travail de thèse.

This study shows that a pre-labeling of word uses by means of a disambiguation algorithm such as *HyperLex* (Véronis, 2003, 2004) allows a better extraction of lexical relations (*NOUN-ADJECTIVE, NOUN "de" NOUN, NOUN-VERB*, etc.), since these relations are categorised with respect to word use. In addition, this technique enables us to retrieve relations for very infrequent word uses, which otherwise would be buried in the residual noise corresponding to the most frequent uses. We performed an evaluation on several thousand web pages containing a target word among a set of 10 highly polysemic ones. We show that the precision obtained is very good, with a quite honourable recall, sufficient in any case for many applications. The analysis of errors opens avenues of research for the rest of our PhD work.

Mots-clefs – Keywords

Corpus, relations lexicales, acquisition automatique, désambiguïsation lexicale.

Corpus, lexical relations, automatic acquisition, word sense disambiguation.

1 Introduction

Le développement d'applications de TAL telles que la traduction automatique ou la recherche d'informations est fortement freiné par le manque de données lexicales détaillées pour chacune des langues. En particulier, les différentes *relations lexicales* (*NOM – ADJECTIF*, *NOM1 DE NOM2*, *NOM – VERBE*, etc.) qui lient le mot à ses cooccurrents en contexte seraient intéressantes à connaître de façon détaillée, car elles sont souvent des indices fortement désambiguïsateurs. Par exemple, pour traduire en anglais le mot *barrage*, il faut au préalable avoir identifié son sens en contexte : *barrage hydraulique* sera traduit par *dam*, *barrage routier* par *roadblock*, *match de barrage* par *play-off game*, etc. On voit dans l'exemple ci-dessus que la simple présence d'un adjectif (*hydraulique*, *routier)* ou d'une relation *NOM1 DE NOM2* (match de *N*), suffit à indiquer le sens du mot dans le contexte incriminé. Il en va de même pour les relations *NOM – VERBE* (sujet, objet) : on *construit* ou on *démolit* un barrage hydraulique, alors qu'on *dresse* ou on *démantèle* un barrage routier.

On ne dispose pas à l'heure actuelle de grande base de données lexicale listant ces « affinités » entre lexèmes du français. La création manuelle d'une telle ressource serait un travail gigantesque. La lente constitution du *Dictionnaire Explicatif et Combinatoire* de (Mel'cuk *et al.*,1984, etc.), qui ne liste qu'une petite partie de ces relations, en est un bon exemple. Or, il semble que si l'on disposait d'une grande quantité de textes informatisés, il pourrait être possible d'en extraire par des moyens automatiques ou semi-automatiques au moins les relations lexicales les plus importantes et récurrentes. Cette recherche de relations lexicales a des points communs avec la recherche de termes complexes, pour lesquels il existe des méthodes et logiciels d'extraction tels que, pour le français, *Acabit* (Daille, 1994) et *Lexter* (Bourigault, 1994). Bien que les relations recherchées ici soient plus générales (toutes ne constituent pas des « termes ») et relèvent des « préférences » lexicales (Wilks, 1975), restrictions de sélection (Katz & Fodor, 1964) ou collocations (Benson, 1990 ; Smadja, 1993 ; Cruse, 1986) générales de la langue, on peut toutefois s'inspirer des mêmes techniques, basées sur la recherche de patrons syntaxiques, comme nous le verrons ci-après. La littérature témoignant d'une terminologie disparate et de concepts souvent flous et contradictoires, nous employons ici le terme *relation lexicale*, plus neutre, défini comme une cooccurrence lexicale entre deux lexèmes liés syntaxiquement. Nous centrons le présent travail sur l'importance d'une catégorisation lexicale automatique des relations lexicales, le tri des relations lexicales extraites sera effectué plus tard.

Une difficulté, lorsqu'on s'intéresse à la langue générale, et non à la terminologie d'un domaine spécialisé, consiste à classifier les relations lexicales obtenues en fonction des différentes acceptions des mots. Le tableau 1 donne par exemple les 10 adjectifs les plus fréquents en cooccurrence avec le mot *barrage* dans le corpus issu du *World Wide Web* utilisé pour cette étude. D'une part, les relations lexicales extraites sont totalement mélangées (on voit qu'il faut ranger les adjectifs *grand*, *hydroélectrique*, etc. avec $barrage_1$ [« barrage hydraulique »]; *routier*, *faux* avec $barrage_2$ [« barrage routier »]). D'autre part, la très grande disparité des fréquences des acceptions (qui suit une loi zipfienne, comme il est connu depuis [Thorndike & Lorge, 1938]) fait que les mots associés aux usages les moins fréquents (par exemple *fatal*, *décisif* associés à *barrage* en tant que rencontre sportive, qui représente moins de 5% des occurrences) sont totalement noyés dans les relations lexicales concernant le ou les usages majoritaires (ici, « barrage hydraulique »).

Les méthodes existantes d'extraction automatique de collocations ne tiennent pas compte des usages des mots. Ainsi, que les auteurs adoptent une approche essentiellement statistique

(Church, Hanks, 1989) ou bien syntaxico-statistique (Smadja, 1993), de nombreuses collocations relatives aux usages peu fréquents des mots sont *a fortiori* omises.

ADJECTIF	Usage	Freq
grand	*hydraulique*	304
petit	*hydraulique*	85
hydroélectrique	*hydraulique*	71
routier	*routier*	68
faux	*routier*	31
mobile	*hydraulique*	27
agricole	*hydraulique*	26
haut	*hydraulique*	23
hydro-électrique	*hydraulique*	22
ancien	*hydraulique*	21

Tableau 1 : Les 10 adjectifs les plus fréquents en cooccurrence avec *barrage*

Nous explorons dans cette communication la possibilité d'exploiter une catégorisation automatique préalable des usages de mots. Toutefois, si une telle catégorisation devait faire intervenir des ressources lexicales importantes (incluant celles que nous cherchons à extraire), nous serions face à une circularité manifeste. (Schütze, 1998) et (Véronis, 2003) ont heureusement montré qu'on pouvait effectuer une catégorisation automatique par l'examen des similarités entre contextes directement à partir de grands corpus sans ressources lexicales (sémantiques) préexistantes. Nous allons plus précisément considérer ici l'aide qu'offre la pré-catégorisation opérée par l'algorithme *HyperLex* (Véronis, 2003, 2004)[1] pour l'extraction des relations de type *NOM – ADJECTIF*, *NOM1 DE NOM2*, *NOM – VERBE*. Cet algorithme permet d'extraire les différents usages d'un mot. L'avantage d'*HyperLex*, contrairement aux méthodes précédemment proposées (vecteurs de mots, cf. Schütze, 1998) est la détection des usages très peu fréquents (de l'ordre de 1% des occurrences).

2 L'algorithme *HyperLex*

Nous exposons très brièvement dans cette section la méthode qui sous-tend l'algorithme *HyperLex* (voir [Véronis, 2003, 2004] pour une présentation détaillée).

A partir d'un mot-cible donné, *barrage* par exemple, seuls les contextes qui contiennent *barrage* sont retenus. Ces contextes sont ensuite lemmatisés et étiquetés morpho-syntaxiquement, et filtrés en fonction d'un certain nombre de critères (élimination des mots fréquents, etc.). Une fois les contextes filtrés, les mots qui apparaissent en cooccurrence avec le mot-cible servent à créer un graphe pondéré par la force d'association entre cooccurrents. Les mots constituent les nœuds du graphe. Les arêtes du graphe représentent les inter-connexions entre les mots qui se retrouvent au sein du même contexte. Ainsi les noms *production* et *électricité* dans l'exemple ci-dessous sont reliés (figure 1).

> *Outre la* ***production*** *d'****électricité****, le* ***BARRAGE*** *permettra de réguler le cours du fleuve…*

[1] Démonstration en ligne : http://www.up.univ-aix.fr/veronis

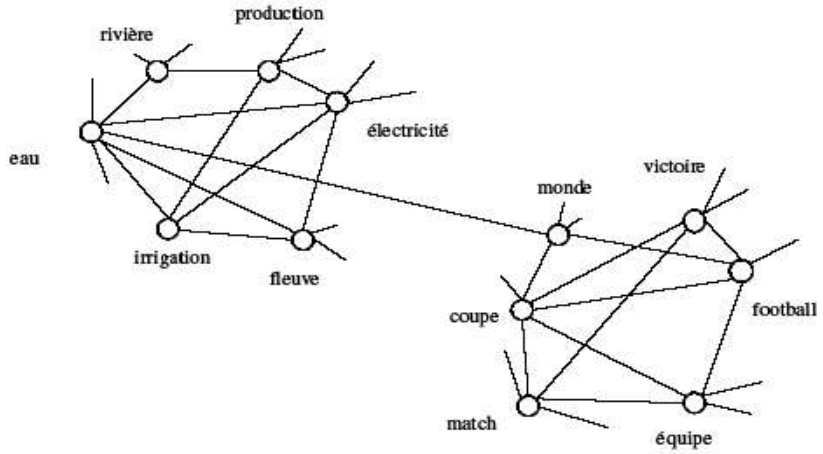

Figure 1 : Graphe des cooccurrences du mot *barrage* (d'après Véronis, 2004)

Les inter-connexions entre les mots font émerger des *composantes fortement connexes* (figure 1), qui correspondent aux différents usages du mot-cible du corpus analysé.

L'algorithme *HyperLex* a été évalué sur 10 mots-cibles très polysémiques (*barrage*, *détention*, *formation*, *lancement*, *organe*, *passage*, *restauration*, *solution*, *station* et *vol*) à partir d'un corpus de plusieurs milliers de pages Web constitué pour chacun des mots-cibles à l'aide d'un méta-moteur de recherche (*Copernic Agent*). Seul un petit nombre d'usages a été omis par l'algorithme : la quasi-totalité des usages de fréquence supérieure ou égale à 5% dans le corpus est correctement détectée, bien que dans quelques cas les regroupements ou absences de regroupements puissent être critiqués. Ainsi pour *barrage*, quatre composantes émergent, que nous désignons par le mot le plus fréquent de chacune d'elles : *EAU*, *ROUTIER*, *FRONTIERE*, *MATCH*, qui représentent bien les emplois dans le corpus recueilli.

Par ailleurs, *HyperLex* fournit un score qui permet d'étiqueter chaque contexte avec l'usage le plus probable du mot-cible qu'il contient (Véronis, 2004). Un coefficient de fiabilité variant entre 0 et 1 est associé à l'étiquetage et permet de régler le rappel. La précision de l'étiquetage est remarquable, puisqu'elle a été évaluée à 97% sur un échantillon aléatoire de 1000 contextes, pour un rappel de 82%, correspondant à un coefficient de fiabilité de 0.5 (l'étiquetage de base ou *baseline*, étant de 73%).

3 Extraction automatique de relations lexicales

Nous reprenons les dix mots-cibles et les corpus de pages Web constitués dans le cadre de l'évaluation de l'algorithme *HyperLex* (voir section précédente). Nous avons écrit un programme qui extrait automatiquement les relations de type *NOM – ADJECTIF*, *NOM1 DE NOM2* et *NOM – VERBE* en fonction de patrons syntaxiques représentés par des expressions régulières simples (tableau 2[1]). Des filtres linguistiques sont ensuite appliqués pour éliminer les candidats adjectivaux, nominaux ou verbaux indésirables, soit parce qu'ils sont trop généraux (noms « quantifieurs », auxiliaires, etc.), soit parce qu'ils correspondent à des erreurs connues et récurrentes dans l'étiquetage morpho-syntaxique.

[1] L'astérisque symbolise les items pouvant se répéter *n* fois ; les parenthèses représentent les items optionnels ; enfin, les items barrés dans les patrons syntaxiques verbaux sont ceux ne pouvant y figurer, le nom et le verbe peuvent être séparés par quatre items textuels.

Pour les noms et les verbes, la place du mot-cible joue un rôle important dans l'identification de son usage en contexte. Le nom *vol* n'a par exemple pas le même sens dans les séquences *carnet de vol* et *vol de carnet.* Deux listes sont ainsi créées pour les verbes et les noms selon la position (antéposée ou postposée) du mot-cible. Les noms propres et les noms communs sont également séparés. Notre programme produit donc pour chaque usage, une liste d'adjectifs, deux listes de verbes et quatre listes de noms.

Nom Adjectif
Adjectif Nom_{CIBLE}
Nom_{CIBLE} (Adverbe)* Adjectif* ((Coordination) (Adverbe)* Adjectif)*
Nom_{CIBLE} (Adverbe)* Adjectif* ((Juxtaposition) (Adverbe)* Adjectif)*
Nom de Nom
Nom_{CIBLE} de Nom2
Nom1 de Nom_{CIBLE}
Nom - Verbe
Nom_{CIBLE} (~~verbe conjugué~~ ou ~~infinitif~~ ou ~~ponctuation forte~~)^{0,4} $Verbe_{\sim\sim ETRE \sim\sim \text{ et } \sim\sim AVOIR \sim\sim}$
$Verbe_{\sim\sim ETRE \sim\sim \text{ et } \sim\sim AVOIR \sim\sim}$ (~~verbe conjugué~~ ou ~~infinitif~~ ou ~~ponctuation forte~~ ou ~~pronoms personnels sujets~~)^{0,4} Nom_{CIBLE}

Tableau 2 : Patrons syntaxiques

Les relations lexicales extraites sont catégorisées automatiquement selon l'étiquetage préalable effectué par *HyperLex*. Aucun filtrage statistique n'est effectué sur les listes obtenues. En effet, nous partageons le point de vue de (Bourigault *et al.*, à paraître) pour qui « l'expérience montre qu'aucune mesure statistique ne peut suppléer l'expertise de l'analyste, en particulier parce qu'il y a toujours des candidats termes de fréquence 1 dont l'analyse est intéressante.» Ce point de vue nous semble encore plus fondé dans le cas de la langue générale, et dans notre cas particulier, où nous essayons de tenir compte des usages très peu fréquents des mots-cibles : la fréquence des usages se répercute tout naturellement sur les fréquences des relations lexicales extraites pour cet usage. De plus, le seuillage appliqué à l'aide du coefficient de fiabilité peut encore réduire les fréquences observées. Ainsi, *barrage de qualification* n'apparaît que 4 fois dans les quelque 7000 contextes de *barrage* dans notre corpus. Après seuillage, il n'en reste qu'une occurrence. Cette relation, bien que de fréquence très faible, est manifestement pertinente pour l'usage *MATCH*.

Etant donné néanmoins que les relations de fréquence 1 (*hapax*) risquent d'être plus fortement bruitées, nous avons considéré les résultats de la validation des relations lexicales de deux manières : dans leur globalité (quelle que soit la fréquence d'apparition des relations lexicales), et en ne tenant compte que de celles de fréquence >1. Nous donnerons donc des résultats concernant quatre configurations : la totalité des relations lexicales extraites au seuil de 0,5 et de 0,6 et sans les hapax aux mêmes seuils.

4 Premiers résultats

Les relations lexicales extraites ont été validées en examinant les contextes d'apparition de leurs occurrences. Il suffit qu'une seule occurrence soit correcte pour que la relation lexicale soit validée. Ainsi la relation *barrage - fermer* de l'usage *EAU* est validée car elle est correcte dans des cas comme :

> *Le* **barrage** *Robert-Bourassa, qui* ***ferme*** *la vallée de La Grande Rivière, est constitué de 23 millions de mètres cubes de remblai.*

alors que l'une des occurrences extraites est incorrecte, le sujet n'étant pas *barrage* mais *bras* :

> *Lors de crues extrêmes, les deux* bras *du* **barrage** *Maesland* ***ferment*** *à la mer du Nord la Nieuwe Waterweg large de 360 mètres.*

Nous ne pouvons, faute de place, présenter des résultats exhaustifs pour chacun des mots. Nous détaillerons donc seulement le cas de *barrage* (et plus particulièrement la relation *ADJECTIF -NOM*) et nous donnerons ensuite des résultats globaux pour les 10 mots-cibles.

4.1 Mot-cible *barrage*

Le tableau 3 donne les 10 premiers adjectifs extraits pour *barrage* avec un coefficient de fiabilité minimum de 0,5, en réponse à ceux présentés de façon non-catégorisée dans l'introduction (tableau 1).

On peut noter que la catégorisation obtenue est très bonne. Seuls deux adjectifs semblent catégorisés de façon incorrecte : une occurrence de *hydraulique* pour l'usage *ROUTIER* et de *hydroélectrique* pour *FRONTIERE*. L'examen des contextes concernés[1] confirme bien qu'il s'agit d'une erreur de catégoristation d'*HyperLex*.

EAU	Freq	ROUTIER	Freq	FRONTIERE	Freq	MATCH	Freq
grand	116	routier	49	terrestre	7	ultime	2
petit	32	filtrant	4	algérien	3	décisif	1
hydro-électrique	29	faux	3	continu	2	fatal	1
mobile	19	fixe	2	frontalier	2	irlandais	1
existant	11	annuel	1	routier	2		
haut	10	différent	1	autoroutier	1		
gonflable	8	grand	1	barbelé	1		
déversant	7	**hydraulique**	1	contemporain	1		
principal	6	illégal	1	**hydroélectrique**	1		
vieux	5	long	1	israélien	1		

Tableau 3 : Les 10 premières relations *barrage - adjectif* au seuil 0,5

Le tableau 4 donne les taux de précision et de rappel (toujours pour le mot-cible *barrage*) pour chacun des types de relations lexicales avec les quatre configurations testées. On voit que la précision est excellente quel que soit le type de relation lexicale, sauf pour la relation *Barrage* (sujet) – *VERBE* dont le meilleur score n'atteint que 78%. La configuration qui a le plus fort taux de précision est la non-considération des hapax au seuil de 0,6, mais avec un taux de rappel assez faible. Il faut noter cependant que notre mesure de rappel est très sévère : nous considérons comme pertinents tous les mots qui entrent en relation dans le corpus avec le mot-cible, sans distinction linguistique aucune. Ainsi, l'absence de *bavarois* dans les adjectifs liés à *barrage* (usage *MATCH*) est considéré comme autant pénalisant que *qualificatif* (tous deux hapax). Or, intuitivement, l'omission du second semble plus dommageable que l'omission du premier, qui ne semble être relié à *barrage* que parce qu'il appartient à une (très grande) classe d'adjectifs géographiques (pays, régions, villes) : ceci est une des directions de recherche future pour notre travail de thèse.

---

[1] Contexte de *barrage - hydraulique* : « Fondée en 1928, la CIBG vise à promouvoir l'art et la science des **barrages** *hydrauliques*. Elle a environ 6 000 membres et des comités dans 80 pays. »

Contexte de *barrage - hydroélectrique* : « Le **barrage** *hydroélectrique* d'Emosson se trouve dans un site spectaculaire surplombant le petit hameau suisse du Châtelard à 19 km de Chamonix. […] »

	TOTALITE				SANS LES HAPAX			
	Précision		Rappel		Précision		Rappel	
seuil	*0,5*	*0,6*	*0,5*	*0,6*	*0,5*	*0,6*	*0,5*	*0,6*
BARRAGE ADJECTIF	0,88	0,91	0,61	0,47	1,00	1,00	0,23	0,16
NOM1 DE BARRAGE	0,98	0,98	0,56	0,41	0,96	1,00	0,15	0,12
BARRAGE DE NOM2	0,91	0,91	0,70	0,56	0,94	0,92	0,26	0,19
VERBE - BARRAGE	0,87	0,90	0,68	0,54	0,97	1,00	0,23	0,17
BARRAGE - VERBE	0,63	0,67	0,72	0,62	0,76	0,78	0,31	0,25

Tableau 4 : Précision et rappel des relations extraites pour *barrage* (tous sens confondus)

Faute de place, nous ne pouvons détailler les résultats par usage pour chacune des relations. Le tableau 5 donne seulement ces résultats pour les adjectifs, en regard de la fréquence estimée (par sondage aléatoire) de chacun des usages. On voit que la précision ne dépend pas de cette fréquence. Notre méthode permet donc d'obtenir des relations lexicales fiables pour des usages peu fréquents des mots-cibles.

	Fréq.	TOTALITE		SANS LES HAPAX	
	estimée	*0,5*	*0,6*	*0,5*	*0,6*
EAU	77%	0,89	0,91	1,00	1,00
FRONTIERE	11%	0,86	1,00	1,00	1,00
ROUTIER	8%	0,80	0,83	1,00	1,00
MATCH	4%	1,00	1,00	1,00	n/d
Total	**100%**	**0,88**	**0,91**	**1,00**	**1,00**

Tableau 5 : Précision des adjectifs extraits pour *barrage*

4.2 Résultats globaux

Le tableau 6 donne les résultats de la validation sur l'ensemble des mots-cibles. On note toujours un fort taux de précision, sauf pour la relation *NOM* (sujet) – *VERBE* (tableau 6). Ces résultats confirment donc les observations faites précédemment sur le mot *barrage*.

	TOTALITE		SANS LES HAPAX	
seuil	*0,5*	*0,6*	*0,5*	*0,6*
NOM - ADJECTIF	0,85	0,88	0,94	0,93
CIBLE de NOM2	0,86	0,89	0,93	0,97
NOM1 de CIBLE	0,92	0,94	0,97	0,97
NOM - VERBE	0,59	0,57	0,70	0,71
VERBE - NOM	0,78	0,80	0,86	0,85

Tableau 6 : Précision globale des relations lexicales validées des 10 mots-cibles

5 Analyse des erreurs et perspectives

Nous avons analysé les erreurs pour l'ensemble des mots-cibles, en les catégorisant selon le type de source qui en est à l'origine : l'étiquetage morpho-syntaxique opéré par *Cordial*, l'étiquetage des usages par *HyperLex* ou l'extraction réalisée par notre programme. Le tableau 7 présente la répartition des erreurs en fonction des trois sources (nous avons traité les relations *NOM de NOM* et *NOM -VERBE* indépendamment de la position).

	Cordial	HyperLex	Extracteur	Total
NOM ADJECTIF	30%	32%	38%	100%
NOM DE NOM	35%	57%	8%	100%
NOM VERBE	20%	12%	67%	100%

Tableau 7 : Taux d'erreurs « Sans les Hapax (seuil 0,5) »

On voit que notre extracteur fonctionne assez bien pour les relations *NOM de NOM*, mais a plus de difficulté avec la relation *NOM-ADJECTIF*, et encore plus avec la relation *NOM-VERBE*, ce qui explique, au moins partiellement, les résultats plus modestes obtenus pour cette catégorie. Nous proposons ci-dessous une analyse un peu plus fine des types d'erreurs.

- Erreurs de *Cordial*

Nom – Verbe : Le logiciel *Cordial Analyseur* produit de nombreuses erreurs d'étiquetage quand un mot a un homographe verbal et qu'il se trouve dans un patron syntaxique analogue à celui *Verbe – Nom* (objet) ou *Nom* (sujet) – *Verbe* comme le nom commun *tonne* et le nom propre *Cuba* dans les attestations suivantes

> *Le second passager d'Ariane était un satellite militaire britannique, Skynet-4E, de 1,5* ***tonne*** *au* **lancement**
>
> *billet d'avion cuba,* **vols *cuba***

L'indice formel qui permet d'identifier un nom propre est la majuscule en initale du mot. Or, dans la plupart des contextes présentant cette particularité, les noms propres sont écrits en minuscules. Une partie de ces erreurs pourra être écartée en ajoutant au protocole de notre extracteur, la congruence flexionnelle des séquences *Nom – Verbe*, le verbe s'accordant avec le sujet. De plus, le logiciel ne reconnaît pas l'auxiliaire du futur proche, à savoir *aller*. Nous pourrions identifier automatiquement quelques-unes des occurrences de l'auxiliaire *aller* grâce à la présence d'un infinitif qui le suit. D'autres erreurs sont liées aux noms *matière*, *cas* et *cours* dans les locutions *en matière de Nom, en cas de Nom, en cours de Nom.* Il suffira de vérifier que ces noms n'entrent pas dans de telles constructions.

- Erreurs d'*HyperLex*

Certaines des erreurs d'*HyperLex* sont dues à la présence de mots fortement connectés à un autre usage du mot-cible que celui qui apparaît dans le contexte. Dans l'exemple suivant, l'occurrence du mot-cible *organe* est rattachée aux « dons d'organes » de façon fautive à cause de la présence du mot (lui-même ambigu) *greffe* :

> *Si l'***organe *de gestion*** *refuse de constater la démission, elle est reçue au* greffe *de la justice de paix du siège social.*

La prise en compte des distances entre mots (en terme de position dans le contexte), ou des relations dans lesquelles ces mots entrent (ici *gestion* est directement relié syntaxiquement à *organe*, contrairement à *greffe*) pourrait amener une amélioration majeure au comportement d'*HyperLex*.

- Erreurs de l'extracteur

Les erreurs de notre programme sont principalement des erreurs de rattachement.

Nom – Adjectif : erreur qui est liée, soit à l'absence de ponctuation (menu de page Web par exemple), soit à l'absence de la construction *Nom1 Adjectif de* Nom_{cible} *Adjectif*:

> pratique officielle ***vivante*** **organe** *consultatif pour étrangers OCEL*
>
> *Mais il n'existe aucun exemple de* rupture *brutale de* **barrage** ***consécutive*** *à un tremblement de terre.*

Cette erreur peut être évitée en ajoutant la séquence *Nom1 Adjectif de* Nom_{cible} *Adjectif* au protocole d'extraction et en vérifiant la congruence flexionnelle entre le mot-cible et l'adjectif candidat.

Nom – Verbe : Dans l'exemple suivant (un autre type de menu de page Web),

> *Emploi* | *Les entreprises qui* ***recrutent*** | **Formation** |,

le mot-cible *formation* n'est pas complément d'objet du verbe *recruter*. Ce type d'erreurs peut être écarté lors du nettoyage des corpus de pages Web. D'autres erreurs découlent des patrons syntaxiques utilisés pour extraire les relations de type *Nom – Verbe.*

> *Le texte* ***crée*** *un* juge *de la* **détention** *provisoire.*
>
> *Les* retenues *des* **barrages** ***régularisent*** *les débits naturels*

Un filtre est appliqué sur ces séquences, à savoir le rejet du verbe candidat si Nom_{cible} entre dans une structure *Nom1 de* Nom_{cible}. Certaines de ces erreurs pourront être évitées facilement en élargissant le filtre *Nom1 de* Nom_{cible} à *Nom1 Préposition* Nom_{cible} (les occurrences du mot-cible ne doivent pas être compléments de *Nom1*) et en vérifiant l'accord entre le verbe et son sujet.

Contrairement aux relations *NOM – ADJECTIF* et *NOM1 DE NOM2*, où les variantes syntaxiques peuvent être facilement appréhendées (l'adjectif épithète par exemple ne peut être séparé du nom qu'il modifie que par des adverbes ou d'autres adjectifs), les relations *NOM – VERBE* (sujet, objet) sont plus délicates à extraire puisque le nom-cible peut être séparé du verbe par des constituants syntaxiques hétérogènes (compléments circonstantielles, relatives par exemple) pouvant être de taille conséquente. De nombreuses relations *Barrage* (sujet) – *VERBE* extraites sont incorrectes. Les récents travaux sur l'analyse syntaxique partielle (shallow parsing) (cf. Abeillé *et al.*, 2003), laissent penser que des progrès significatifs pourraient être fait sur l'extraction de ce type de relations.

6 Conclusion

Nous avons essayé de montrer dans cet article qu'un pré-étiquetage des usages des mots par un algorithme tel qu'*HyperLex* (Véronis, 2003, 2004) permettait d'obtenir des relations lexicales (du type *NOM-ADJECTIF*, *NOM de NOM*, *NOM-VERBE*) beaucoup plus exploitables, parce qu'elles-mêmes catégorisées en fonction des usages. De plus, cette technique permet d'obtenir des relations pour des usages très peu fréquents, alors qu'une extraction indifférenciée « noierait » ces relations au milieu de celles correspondant aux usages les plus fréquents. Nous avons conduit une évaluation sur un corpus de plusieurs milliers de pages Web comportant l'un des 10 mots-cibles très polysémiques choisis pour cette expérience, et nous avons montré que la précision obtenue était très bonne, avec un rappel

honorable, suffisant en tout cas pour de nombreuses applications. L'analyse des erreurs nous a montré des perspectives d'améliorations pour la suite de notre travail de thèse. Nous étendrons notre étude à d'autres mots du français. La base de données lexicale que nous projetons de réaliser ainsi que les nombreuses attestations en corpus que nous aurons relevées permettront d'établir une typologie plus précise des phénomènes lexicaux, et de clarifier le débat sur la relation entre items lexicaux et leur possibilités et contraintes combinatoires.

Références

Abeillé A., Clément L., Toussenel F. (2003), Building a treebank for French, *Abeillé A. Treebanks. Building and unising parsed corpora,* pp. 165-187.

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

Bourigault D. (1994), LEXTER, un Logiciel d'EXtraction de TERminologie. Application a l'acquisition des connaissances à partir de textes, Ph.D. Thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.

Bourigault D., Aussenac-Gilles N., Charlet J. (à paraître), Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, *Revue d'Intelligence Artificielle* [En ligne : http://www.univ-lse2.fr/erss/membres/bourigault/]

Church K., Hanks P. (1989), Word association norms, mutual information, and lexicography, *Computational Linguistics*, Vol. 16(1), pp. 76-83.

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

Daille B. (1994), *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques,* Paris : Université de Paris VII.

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

Mel'cuk I. A., Arbatchewsky-Jumarie, N., Elnitsky, L., Iordanskaja, L., Lessard. A. (1984), *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I,* Montréal, Presses de l'Université de Montréal.

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

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

Thorndike E. L., Lorge I. (1938), *Semantic counts of English Words,* New York, Columbia University Press.

Schütze H. (1998), Automatic word sense discrimination, *Computational Linguistics*, Vol. 24 (1), pp. 97-124.

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

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