RÉCITAL 2002, Nancy, 24-27 juin 2002

Méthodologie pour la création d'un dictionnaire distributionnel dans une perspective d'étiquetage lexical semi-automatique

Delphine Reymond

Équipe DELIC – Université de Provence
29, Av. Robert Schuman, 13621 Aix-en-Provence
Delphine.Reymond@up.univ-aix.fr

Mots-clefs – Keywords

Désambiguïsation lexicale, dictionnaire, propriétés distributionnelles, collocations, classes d'objets

WSD, dictionary, distributional properties, collocations, classes of objects

Résumé – Abstract

Des groupes de recherche de plus en plus nombreux s'intéressent à l'étiquetage lexical ou la désambiguïsation du sens. La tendance actuelle est à l'exploitation de très grands corpus de textes qui, grâce à l'utilisation d'outils lexicographiques appropriés, peuvent fournir un ensemble de données initiales aux systèmes. A leur tour ces systèmes peuvent être utilisés pour extraire plus d'informations des corpus, qui peuvent ensuite être réinjectées dans les systèmes, dans un processus récursif. Dans cet article, nous présentons une méthodologie qui aborde la résolution de l'ambiguïté lexicale comme le résultat de l'interaction de divers indices repérables de manière semi-automatique au niveau syntaxique (valence), sémantique (collocations, classes d'objets) avec la mise en œuvre de tests manuels.

More and more research groups are involved in sense tagging or sense disambiguation. The current trend is to use very large text corpora which, with the help of appropriate lexicographical tools, can provide initial data to the disambiguation systems. In turn, these systems can be used to extract more data from corpora, which can be fed again to the systems, in a bootstrapping process. In this paper, we tackle lexical disambiguation through the interaction of various cues which can be detected semi-automatically at the syntactic and semantic levels (valency, collocations, object classes), along with manual tests.

1 Introduction

Ce travail de thèse part d'une réflexion sur le problème de la désambiguïsation lexicale, problème fondamental dans le domaine de la linguistique computationnelle (Ide & Véronis, 1998). Le traitement de la polysémie étant particulièrement délicat, on constate malheureusement qu'à l'inverse des systèmes d'étiquetage morpho-syntaxique qui ne cessent de s'améliorer, les systèmes d'étiquetage lexical ne progressent guère. La désambiguïsation consiste à déterminer la signification qu'un mot peut avoir dans un contexte particulier par rapport à la liste des sens d'un dictionnaire. Certains travaux (Véronis, 2001) ont démontré que les dictionnaires usuels tels le *Petit Larousse* sont inadéquats quand il s'agit d'étiqueter le sens des mots. En effet, la performance « humaine » sur l'identification des « sens » est particulièrement médiocre. Le bilan n'est guère encourageant car il est difficile d'espérer de meilleurs résultats par un traitement automatique, si la performance humaine n'est que peu concluante pour ce genre de tâche.

Notre attention s'est portée vers le développement de connaissances lexicales à partir de grands corpus textuels pour la constitution d'un dictionnaire distributionnel. Dans cette perspective, nous pensons qu'il est plus intéressant de lister un mot et ses usages possibles en contexte plutôt que ses différentes acceptions. Cette approche n'est pas nouvelle puisque dès les années 20, Meillet affirmait que « le sens d'un mot ne se laisse définir que par une moyenne entre [ses] emplois linguistiques ». Wittgenstein prenait en 1953 une position analogue (« Do not look for the meaning, but for the use »). Cette vision distributionnaliste n'a pourtant été implémentée que de façon très partielle en lexicographie.

La lexicologie anglaise a été la première à fournir des ressources linguistiques offrant une approche distributionnelle de la langue à partir de l'examen systématique de corpus informatisés : le LDOCE[1] et l'OALD[2], par exemple, incluent des codes syntaxiques et sémantiques décrivant de façon grossière le fonctionnement des lexies, et leurs versions électroniques ont largement été utilisées pour le T.A.L.. Pour le français, les ressources lexicales informatisées relevant l'usage des mots dans les corpus sont à peu près inexistantes, même si la récente informatisation du *Trésor de la Langue Française* laisse entrevoir des perspectives d'exploitation. Les dictionnaires de spécialités décrivant l'usage en corpus commencent à être publiés, comme le DAFA[3] (Binon & al, 1999), basé sur un corpus de 25 millions de mots, qui présente quelques 3200 mots avec leurs contextes d'emploi, collocations et où chaque article décrit un champ sémantique (synonymes, antonymes, hyperonymes, etc).

Notre méthodologie s'inspire en partie du dictionnaire explicatif et combinatoire de Mel'Cuk (1995) où les propriétés distributionnelles permettent de formaliser la distribution d'un prédicat (ou lexie) dans une phrase considérée comme représentative de l'usage en contexte. La démarche est cependant différente puisque le corpus est utilisé comme base empirique et non comme outil de vérification. De plus, notre approche est purement discriminante et non définitoire. Le corpus et le dictionnaire interagissent : le premier est l'outil de référence pour

[1] Longman Dictionary of Contemporary English

[2] Oxford Advanced Learner's Dictionary

[3] Dictionnaire d'Apprentissage du Français des Affaires

l'élaboration du dictionnaire qui a son tour va devenir l'outil de référence pour l'élaboration de corpus annotés. Nous essayons de constituer un dictionnaire basé sur un ensemble de critères différentiels les plus stricts possibles, codés dans l'entrée lexicale et pouvant constituer ultérieurement autant d'indices exploitables par des machines pour discriminer le sens des mots en contexte (Reymond, 2001). A ce stade, notre travail se présente sous la forme d'un dictionnaire distributionnel de 60 vocables polysémiques (20 noms, 20 verbes et 20 adjectifs) où chaque occurrence (environ 60 000 au total) a été étiquetée manuellement dans un corpus d'un million de mots.

2 Typologie des indices

2.1 Valence

La notion de valence est primordiale puisqu'elle est fondée sur les restrictions sélectionnelles entre un prédicat et ses dépendants. De nombreux travaux se sont penchés sur l'étude de la valence verbale comme le projet Proton (Van den Eynde & Mertens, 2001) qui utilise les propriétés discriminatoires des paradigmes de pronoms. Dans notre approche, nous limitons la description du régime syntaxique d'une lexie au nombre d' « actants » qu'elle accepte sans distinction précise entre actants, circonstants ou adjonctions (ex : X_{qqn} *donne* Y_{qqch} *à* Z_{qqn}).

Les indices syntaxiques sont porteurs d'informations lexicales et ils sont en partie automatiquement détectables par une analyse partielle (« *shallow parsing* »). A l'aide d'un outil élaboré de concordance (CooLox[4]) sur corpus étiqueté morpho-syntaxiquement, nous pouvons décrire une cible et ses contextes par une méta-expression régulière où les propriétés morpho-syntaxiques des mots du corpus peuvent être intégrées dans la requête (Audibert, 2001).

Prenons l'exemple des constructions [verbe + préposition] pour lesquelles le corpus de référence[5] préalablement étiqueté par *Cordial 7 Analyseur*[6], offre un grand nombre de contextes d'emplois différents (*répondre de qqch/qqn, passer par, en venir à,* etc.). La figure 1 illustre la requête qui permet d'extraire toutes les lignes de concordance dont la cible est une occurrence dont l'étiquette morpho-syntaxique (EMS) commence par « V » (verbe), suivie d'une occurrence dont l'étiquette commence par « PREP » (préposition). On obtient l'affichage de la figure 2.

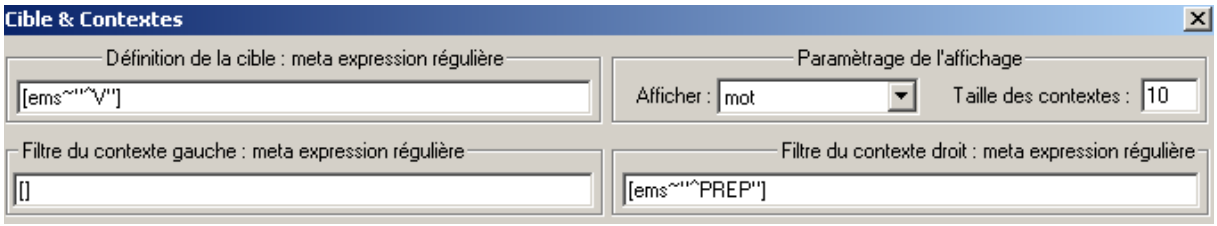

Figure 1 : Requête sur les propriétés morpho-syntaxiques des mots [V+PREP]

[4] CooLox disponible sur http://laurent.audibert.free.fr/Pages/Recherche/Lox/Lox.htm

[5] Corpus SyntSem, d'environ 5 millions de mots, constitué d'œuvres littéraires (A), de textes journalistiques (M), de périodiques (P) et d'ouvrages scientifiques (O).

[6] Cordial 7 Analyseur. © 2000 Synapse Développement. http://www.synapse-fr.com/

Contexte gauche	Cible	Contexte droit
' adressant à la foule : « La Lune ,	prenant	en pitié le souverain cher aux enfants de l '
jeune femme sur les deux joues , en la	prenant	par les épaules , ce qui chiffonna un peu l
on des ballons , ont tiré dessus , les	prenant	pour de monstres aériens ; il est donc permis à
êmes . On dirait vraiment qu ' il nous	prend	pour ses lecteurs . " Puis ils descendirent sur l
une petite cassette . Eh bien , j ' en	prendrai	pour vingt mille ! Tu te mets de moitié .
; et il fut convenu que ses témoins le	prendraient	chez lui en landau , le lendemain à sept heures
ressemblent à des îles , ceux qu ' on	prendrait	pour des montagnes de neige - - tâchant de distin
il réfléchit à la façon dont il s ' y	prendrait	pour se procurer le repas du soir . A sept
s donne pas six mois pour vous laisser	prendre	à cet appât - là . Vous serez madame la
Nil ! répéta Kennedy , qui se laissait	prendre	à l ' enthousiasme de Samuel Fergusson . - -
obliquement , les yeux noyés , faisant	prendre	à sa figure une expression mystique . On entendit
. Mme Forestier ne se laisserait point	prendre	à ses adresses , et il perdrait par sa couardise
e lui plut tellement qu ' il désira la	prendre	à son service pour aider la vieille Germaine . Pé

Figure 2 : Lignes de concordances extraites

La recherche peut être affinée par l'écriture d'une méta-expression régulière combinant plusieurs propriétés des mots (mot, lemme, EMS, etc.) comme dans l'exemple de la figure 3 : on affiche les lignes de concordances dont la cible est une occurrence du lemme *passer* suivie d'une occurrence du lemme *pour* suivie d'un nom qui apparaît dans un contexte de 0 à 3 mots maximum après la préposition.

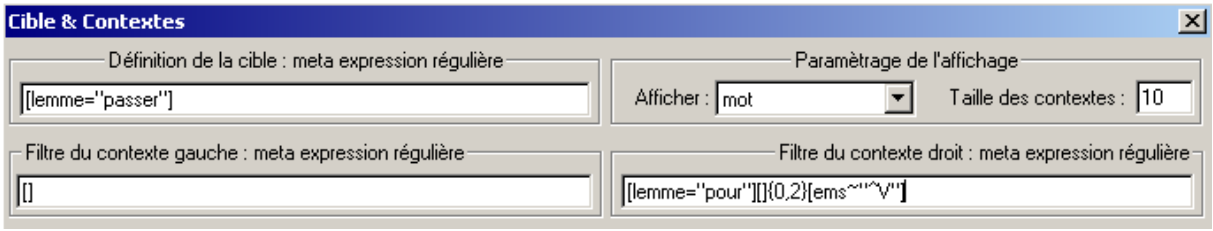

Figure 3 : Requête sur les lemmes et les propriétés morpho-syntaxiques des mots

L'utilisation de ce type d'outil lexicographique permet d'extraire par la simple description de patrons distributionnels toutes les occurrences d'une lexie « candidate » figurant dans le dictionnaire. Après une vérification manuelle, chaque occurrence peut recevoir l'étiquette sémantique appropriée de manière automatique en la rajoutant dans la colonne insérée après l'occurrence cible (figure 4). Le corpus est donc progressivement enrichi selon un mécanisme itératif.

Contexte gauche	Cible	Etiquette	Contexte droit
La manière de montrer ne peut en effet	passer	1.4.1	pour innocente . L ' hype
et de sociologie pourraient facilement	passer	1.4.1	pour la parodie d ' un sty
vre d ' un auteur connu . Elle déteste	passer	1.4.1	pour jalouse ! Au risque
, De guerre lasse , qui n ' a pas	passé	1.4.1	pour un échec . De cela ,
était très belle , très charmante , et	passant	1.4.1	pour une femme " indépend
mpêcher les citrus malades de se faire	passer	1.4.2	pour des yuccas et le pat
tent des Beurs . On a voulu nous faire	passer	1.4.2	pour des terroristes . "
salut . Grâce à lui , je décidai de me	passer	2.5	pour commencer du Figaro

Figure 4 : Exemple d'étiquetage lexical manuel

2.2 Combinaisons de mots

2.2.1 Cooccurrence, collocation et figement

Paradoxalement, la langue est « régulièrement » complexe. La complexité s'apparente au fait que les mots peuvent se combiner entre eux de façon très libre, ce qui rend difficile toute systématisation descriptive. A l'inverse, on note pourtant des affinités de mots les uns pour les autres qui rendent prédictibles certains schémas de construction. Ces affinités ont reçu diverses dénominations. Igor Mel'Cuk parle de *phrasème* ou *semi-phrasème* que Bernard Pottier décrirait plutôt comme *lexie complexe* ; on note également les appellations *expression idiomatique*, *locution* et bien entendu *collocation*. Chez Gaston Gross (1996), le phénomène est appelé *figement*. Les collocations figurent dans la plupart des dictionnaires traditionnels mais on les retrouve généralement dans les exemples qui sont peu nombreux. D'un autre côté, les dictionnaires de collocations se développent, tel le *Dictionary of English Word Combinations* (18 000 entrées et 90 000 collocations) qui offre une excellente description du fonctionnement de la langue anglaise.

La frontière entre restriction, collocation et (semi-)figement est extrêmement difficile à établir. Une combinaison semi-figée de mots n'est-elle pas une collocation ? Parfois, un groupe de mots en collocation forme une locution plus ou moins figée. Katz et Fodor (1964) se réfèrent à des restrictions de sélection de type booléen pour contraindre l'interprétation sémantique. Par exemple, le mot « bordeaux » désigne aussi bien une couleur que du vin. Dans le contexte « Marie boit du bordeaux » le sens de vin est par élimination le seul disponible car le verbe *boire* implique une restriction sur l'objet direct décrite par la classe des <*liquides*>. Les exemples suivants tirés du corpus, montrent que le choix du sens approprié d'un mot en contexte peut être sémantiquement restreint par les éléments en cooccurrence :

1) *réparer, remettre à neuf les organes* (d'un véhicule) mais **réparer, remettre à neuf les organes sexuels*
2) *transplantation, don d'organes* mais **transplantation, don, trafic d'organes de traction / de roulement*

Notre attention s'est donc principalement tournée vers l'étude des cooccurrences. Le corpus de référence étant étiqueté morpho-syntaxiquement, nous avons pu dénombrer dans un contexte maximum de trois mots à gauche et à droite, quelles sont les cooccurrences de mots les plus fréquentes entre les principales parties du discours. Par exemple, le vocable *concentration* possède environ 246 occurrences dans le corpus. La combinaison N_1 de N_2 la plus fréquente est *camp de concentration* (11 occurrences soit 4%), qui est également le premier niveau de sens pour l'entrée.

Ce premier repérage de « surface » peut alors être complété par un travail manuel de filtrage visant à distinguer parmi les combinaisons les plus fréquentes celles qui sont purement « accidentelles », collocationnelles ou figées. Quand il s'agit de collocations, nous avons veillé à ce qu'elles soient également triées en fonction de la catégorie grammaticale et de la place des éléments régis ou recteurs.

▼	**Adj**	***Concentration** verticale*
		***Formation** continue*
	SPrep	***Détention** d'armes*
		***Compagnie** des eaux, de chemin de fer*
▲	**N**	*Opération de **concentration***
		*Condition, lieu, centre de **détention***
	V(O)	*Amender, réviser, modifier, adopter, rédiger la **constitution***

Tableau 1 : tableau des collocations avec les éléments régis (▼) et recteurs (▲)

Nous estimons que la cooccurrence est un degré plus ou moins fort de restriction et qu'elle se situe à plusieurs niveaux (lexical ou syntaxique), le plus bas étant la collocation et le plus élevé étant le figement. Nous distinguerons et étudierons les phénomènes suivants :

1. Les **figements** ou semi-figements dès lors que l'on observe un blocage syntactico-sémantique (ex : *au pied de la lettre, conduire le cotillon*, etc).

2. Les **collocations grammaticales** pour lesquelles on observe un lien entre un mot et une préposition (*répondre de qqch, se mettre à qqch*) ou une construction grammaticale (*il ne se passe pas x temps sans que, il est courant de*).

3. Les **collocations lexicales** en majorité représentées par des combinaisons binaires entre les principales parties du discours et dont voici quelques exemples tirés du corpus :

   V-N : *Couvrir une zone, arrêter une date, conduire une voiture, exercer des pressions*
   ADJ-N : *horloge biologique, effets secondaires, communication politique, concentration verticale, eau courante, air frais, haute antiquité*
   N_1-N_2 ou **N_1 de N_2 :** *Chef-adjoint, gouverneur-chef, barrage-réservoir, station-service, chef de famille / de file, formation du personnel*
   ADV-ADJ : *fort utile*
   ADV-V ou **V-ADV :** *bien se connaître, se mettre debout*
   V-ADJ : *passer inaperçu*

L'analyse quantitative des mots du voisinage est une méthode statistique couramment utilisée. Daille (1994) s'est intéressée entre autres aux relevés des fréquences pour les constructions N_1 de N_2. Sur ce modèle, nous pouvons fournir les fréquences d'association pour les couples [verbe-objet]. L'exemple suivant illustre une extraction des cooccurrences nominales apparaissant dans une fenêtre de 3 mots à droite du lemme *ouvrir* triées sur la fréquence. Toutes les données ne figurent pas dans le tableau.

Lemmes	Fréquence absolue
ouvrir - porte	36
ouvrir - oeil	20
ouvrir - bouche	11
ouvrir - fenêtre	11
ouvrir - bras	5
ouvrir de grand oeil	4
ouvrir - portière	4
ouvrir - porte	4
…	…
ouvrir - boutique	1
ouvrir - passage	1

Tableau 2 : dénombrement des constructions $V_{(ouvrir)}$-N dans le corpus SyntSem

Certaines paires de mots (*ouvrir un passage*) ne sont pas très fréquentes bien que très familières, ce qui soulève le problème de la taille et de la représentativité du corpus, que nous n'aborderons pas ici.

2.2.2 Traits et classes d'objets

Selon la plupart des grammairiens, les phrases sont constituées d'opérateurs qui se déterminent par la nature des arguments qu'il peuvent recevoir. Certaines de nos lexies sont subdivisées en traits syntactico-sémantiques qui permettent de préciser les qualités de l'opérateur. Les traits les plus utilisés sont : animé, non animé, concret, abstrait, etc. On peut par exemple, opposer *un organe politique* à *un organe femelle* où le premier emploi décrit la cible comme étant *animé humain* et le second comme étant *concret*. Cette classification en trait est acceptable mais dans certains cas, elle s'avère insuffisante surtout en vue d'un encodage ou d'un décodage automatique. Considérons les deux exemples suivants :

(M) *La police arrête plusieurs Thai à Anvers, dont Santi.*

(A) *Sa majesté lui arracha des mains le citron […] quand l'archevêque l'arrêtant lui dit à l'oreille :…*

où les sujets et objets reçoivent la caractéristique *animé humain* mais où les lexies semblent distinctes. L'application d'un simple test manuel comme la dérivation du verbe en nom confirme cette supposition : *l'arrestation de qqn par qqn* mais *l'arrêt de qqn par qqn*.

L'alternative que nous avons choisie est celle proposée par Gaston Gross (1997) et le LLI[7] qui préconise l'emploi de classes d'objets (Le Pesant & Mathieu-colas, 1998) dont les divers éléments sont caractérisés par les *prédicats appropriés.* La tâche consiste alors à dresser la liste de tous les membres d'une classe d'objets en position sujet ou objet et par la même occasion d'en décrire le comportement syntaxique.

7 Laboratoire de Linguistique Informatique

Porter, V	**Exemple du corpus**
<*marque informative*>	*Porter une signature, la mention, une inscription*
<*dénomination*>	*Porter le nom, le surnom, le titre (de)*
<*vêtement, accessoire*>	*Porter un chapeau, des lunettes, un manteau, une robe, un uniforme*
<*foetus*>	*La femelle porte les petits, elle porte un enfant*
<*partie du corps*> + Adv	*Porter haut la tête, la tête basse, le chef branlant*
<*Instrument, arme*> qui porte	*Clairon, trompe qui porte haut / loin*
<*compliment, sentiment*> qui porte	*Argument qui porte* mais **argument qui porte haut / loin*
<*phénomène climatique*> qui porte	*Vent, force, courant qui porte quelque part*
Station, N	**Exemple du corpus**
Station <*poste émetteur*>	*Station radiophonique, de télévision, d'émission*
Station <*de recherches, d'observations*>	*Station géophysique, spatiale, sismologique*
Station <*lieu d'activité touristique*>	*Station balnéaire, de ski, de campagne*
Station <*lieu d'arrêt pour transport*>	*Station de métro, de train*
Station <*ressources, services*>	*Station service, d'épuration, de pompage*
Station <*position du corps*>	*La station debout*

Tableau 3 : utilisation des classes d'objets pour la description de lexies

3 Autres critères différentiels et tests manuels

Les limites de l'automatisation des indices de surface marquent le champ de travail qu'il reste à effectuer manuellement pour mener à bien la représentation et le classement des unités lexicales.

a) Sur les 20 adjectifs traités, très peu d'emplois sont identifiables par un comportement syntaxique plus ou moins figé (ex : *être utile à, il est courant de*). Le tableau 4 présente l'entrée simplifiée de l 'adjectif *frais* pour lequel les lexies sont décrites en fonction des synonymes et antonymes substituables sur l'axe paradigmatique.

Frais, Adj.	**Exemple du corpus**
≈ froid	*Vent frais, zone fraîche*
↯ pollué	*Air frais*
↯ vieux	*Traces fraîches, herbe fraîche*
≈ pur	*Voies fraîches*
↯ En conserve, rassis, avarié	*Viande fraîche, poisson, légume, œuf, produit frais*
≈ Reposé, en forme↯ fatigué	*Être frais et dispos*

Tableau 4 : synonyme approximatif (≈) et antonyme approximatif (↯)

b) La position de l'épithète (antéposé ou postposé au nom) nous semblait être potentiellement significative, mais nous n'avons pas relevé de cas suffisamment convaincants qui pourrait attester de la pertinence de la distribution de l'adjectif autour du noyau nominal.

Une salle, poche, voix pleine (? *une pleine salle, poche, voix*)
Population, cellule saine (? *une saine cellule, population)*
Une simple coïncidence, constatation, fait (? *une coïncidence simple, un fait simple), une pleine lune* (? *une lune pleine)*

c) Les tests portant sur la dérivation différentielle avec conservation de la valence, nominalisation quand il s'agit d'un verbe ou verbalisation dans le cas d'un nom, sont relativement efficaces.

(M) *Il fallut tirer à la mitraillette lourde*⇨*le tir à la mitraillette lourde* mais **le tirage à la mitraillette lourde*
(A) *Je vous apporterai un livre et vous le tirerez à 5000.* ⇨*le tirage à 5000* mais **le tir à 5000*

d) Nous avons introduit la notion de classes d'objets pour l'identification des usages et le classement des cooccurrences. Parallèlement, nous avons observé que lorsque plusieurs occurrences acceptent un même hyperonyme c'est qu'elles appartiennent à une même classe d'emploi (figure 5). A l'inverse, l'absence d'hyperonyme peut être également considérée comme différentielle (figure 6).

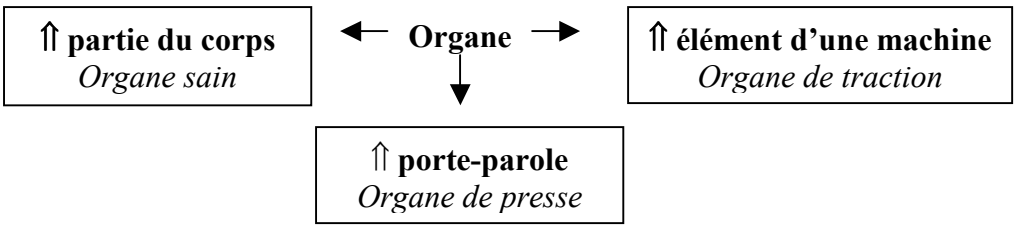

Figure 5 : Découpage et description du vocable organe par l'hyperonyme

Figure 6

4 Conclusion

Notre approche vise à repérer des informations distributionnelles à partir de réalisations rencontrées sur corpus en vue de la constitution interactive de dictionnaires pour le T.A.L. et pour l'annotation sémantique de grands corpus textuels. Nous avons observé que de nombreux indices de « surface » pouvaient être utilisés de façon semi-automatique pour la description des patrons distributionnels des lexies. A l'aide d'un outil élaboré de concordance sur corpus morpho-syntaxiquement étiqueté, nous utilisons ces indices par couches successives dans un processus d'étiquetage manuel « vertical », où toutes les occurrences de distribution analogue sont étiquetées en bloc. Le processus est itératif, puisque l'information ainsi rajoutée au corpus peut être à nouveau utilisée pour l'étiquetage des autres lexies. Il est évident que l'automatisation d'un tel processus ne peut être que partielle, et qu'une intervention manuelle importante est nécessaire. Nous récoltons actuellement les résultats d'une série de tests de performance d'un système de désambiguïsation automatique utilisant l'information codée dans nos entrées (Audibert, 2002). Nous pensons que celles-ci semblent décrire mieux l'usage réel des mots que les dictionnaires classiques dans une perspective de T.A.L.

Références

Audibert, L. (2002). Étude des critères de désambiguïsation sémantique automatique : présentation et premiers résultats sur les cooccurrences. Actes de la Conférence *Traitement Automatique des Langues (RECITAL'2002)*. Nancy (France). A paraître.

Audibert, L. (2001). LoX : Un outil polyvalent pour l'exploration de corpus annotés. Actes de la Conférence *Traitement Automatique des Langues (RECITAL'2001)* (pp. 411-419). Tours (France): ATALA. http://www.up.univ-mrs.fr/delic/papiers/Audibert-2001recital.pdf

Binon, J., Verlinde, S., Van Dyck, J., & Bertels, A. (1999). *Dictionnaire d'apprentissage du français des affaires*. Paris: Didier. 650 pp.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de Doctorat en Informatique Fondamentale. Université Paris 7. 1994. daille_these94.ps.gz

Gross, G. & Clas, A. 1997. «Synonymie, polysémie et classes d'objets», *Meta,* 42(1) Presses de l'Université de Montréal, pp. 147-155.

Gross, Gaston. 1996. *Les expressions figées en français : noms composés et autres locutions.* Paris : Ophrys. 161 p.

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics, 24*(1), 1-40. http://www.up.univ-mrs.fr/~veronis/pdf/1998wsd.pdf

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, pages 479--518. Prentice Hall.

Le Pesant, D., Mathieu-Colas, M. (eds.). 1998. *Les classes d'objets. Langages*, 131. Paris : Larousse.

Mel'cuk, I., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Louvain-la-Neuve : Editions Duculot (Coll. Universités Francophones).

Reymond, D. (2001, 2-5 juillet). Dictionnaires distributionnels et étiquetage lexical de corpus, Actes de la Conférence Traitement Automatique des Langues (RECITAL'2001) (pp. 479-488). Tours (France): ATALA. http://www.up.univ-mrs.fr/delic/papiers/Reymond-2001recital.pdf

Van den Eynde, K. & Mertens, P. (2001, submitted for publication) *La syntaxe du verbe, l'approche pronominale et le lexique de valence PROTON* Preprint 174, Departement of Linguistics, K.U.Leuven, pp. 36.

Véronis, J. (2001). *Sense tagging: does it make sense?* Paper presented at the Corpus Linguistics'2001 Conference, Lancaster, U.K. http://www.up.univ-mrs.fr/~veronis/pdf/2001-lancaster-sense.pdf