Nature et formation de classes sémantiques de verbes pour l'extraction de connaissances dans des textes:

Esquisse d'une approche statistico-symbolique.

Alda Mari (1), Patrick Saint-Dizier (2)

(1) EHESS, 9, rue de Mezieres, 75006 Paris France Alda.Mari@ehess.fr
(2) IRIT - CNRS, 118 route de Narbonne, 31062 Toulouse France stdizier@irit.fr

Abstract

In this document we present a method for classifying verb-senses and for associating with each class, of an appropriate granularity, a set of symbolic and statistical elements that allow for an efficient and accurate knowledge extraction procedure from various types of texts. We concentrate here on predicate-argument structures and introduce a method that combines statistical observations with symbolic descriptions.

Mots clés: Sémantique lexicale, extraction de connaissances dans des textes, formation de classes de verbes

1. Problématique

Notre objectif sur le long terme est l' extraction de connaissances dans des textes, a priori de tout type, mais plutôt ' structurés' tels que des textes scientifiques ou journalistiques. Nous nous intéressons en particulier aux événements relatés dans ces textes, plutôt qu' aux informations `descriptives' . Ces dernières sont en effet souvent représentées par des structures nominales alors que les événements sont essentiellement pris en compte par des prédicats, dont en grande partie des verbes. Notre objectif est donc de pouvoir extraire des structures verbe-arguments dans des textes et de leur associer une représentation sémantique, relativement superficielle, mais qui va nettement au-delà de mot-clés, même structurés.

Pour réaliser cet objectif, il faut dans un premier temps mener une réflexion sur la structure des formes prédicatives, dans le cadre d' une théorie du sens ou de la représentation du sens qui soit adéquate par rapport à nos objectifs. Les exigences de l' extraction de connaissances, dans l' état actuel de l' art, nous conduisent à tenter de regrouper les verbes en petites familles ayant des sens et des réalisations linguistiques apparentées. Le but est d' homogénéiser l' extraction des arguments et la représentation du sens produite. En complément, et au préalable, il faut adopter une stratégie pour délimiter les `sens' d' unités lexicales polysémiques.

Le travail présenté ici est effectué sur le français et s' appuie sur plusieurs travaux antérieurs menés dans notre laboratoire sur l' analyse et la formation de classes sémantiques de verbes. Notre contribution dans ce document consiste en

(1) une prise en compte fine par les biais symboliques et statistiques de la variété des réalisations linguistiques d' un sens de verbe donné (en particulier les formes métonymiques et métaphoriques) afin d' améliorer l' identification des sende ' base' rencontrés dans les textes,
(2) une première évaluation de l' intégration symbolique-statistique, loin d' être un phénomène simple,

(3) l' introduction de critères permettant de prédire un certain ombre d' informations sémantiques sur ce verbe, au vu d' emplois d' un verbe, ou inversement de prévoir l' expansion des classes formées.

2. Formation de classes de verbes

Dans cette section, nous présentons brièvement les techniques de constitution de nos classes et leurs justifications, nous montrons ensuite comment ces classes sont enrichies (1) au niveau des réalisations en langue par une forme simple d' apprentissage statistique à partir de corpus marqués et (2) par le biais de représentations symboliques. Nous analysons enfin l' intégration de ces deux approches complémentaires.

2.1 Constitution de classes sémantiques de verbes

Il aurait pu paraître évident d' utiliser WordNet (ou EuroWordNet) comme système de classification. Toutefois, nous n' avons pu utiliser WordNet directement pour plusieurs raisons: la distinction des sens y est peu claire, un usage correspondant souvent à un sens autonome particulier. Par exemple le verbe ' cut' a environ 25 sens, alors que selon notre approche nous en dégageons 4, tout en ayant la même couverture au niveau des emplois (Gayral et al. 99). Par ailleurs, dans WordNet, il n' y a pas de lien explicite à la syntaxe, ce qui est une limitation forte en extraction de connaissances où l' on a besoin de retrouver les arguments d' un prédicat quelque soit leur réalisation syntaxique. En effet, les seuls éléments syntaxiques présents dans WordNet sont des formes simplifiées de schémas très généraux de sous-catégorisation associés à des restrictions de sélections relativement pointues. Rien de particulier n' est précisé sur les différentes formes syntaxiques (par exemple des alternances) que peuvent prendre les prédicats. Ceci n' est pas une critique en soi de WordNet, dont la description syntaxique n' était pas le but. Enfin, WordNet n' est pas associé à des données quantitatives sur les réalisations incluant métaphores et métonymies qui sont souvent traitées comme des sens à part entière.

Notre démarche a été la suivante (Mari et Saint-Dizier 97), (Saint-Dizier 98): nous avons répertorié les alternances, dans le style de (Levin 93) les plus usuelles pour le français et avons associé à chacun des 1700 verbes les plus courants du français la liste des alternances qu' il accepte.

Notre travail a porté sur des sens de verbes, dans une perspective sémantique générative, avec une vision des sens assez larges. Un sens est distingué des autres sens du même lexème par une grille thématique (fine) qui lui est associée. C' est une dimension qui n' est pas prise en compte a priori de façon systématique dans (Levin 93).

Il est ensuite simple de former des classes d' équivalence à l' intérieur desquelles tous les verbes acceptent exactement les mêmes alternances. A partir de 1700 verbes, nous avons obtenu 953 classes (en accord sur ce point avec les observations de M. Gross, bien que dans un cadre différent). Sur ces 953 classes, 741 comptent un seul élément, 123 ont 2 éléments et 31 ont 3 éléments. Ce qui fait que 56% des verbes sont dans des classes d' au moins 2 éléments, et 32.6% dans des classes ont au moins 5 éléments.

Nos travaux diffèrent de ceux de M. Gross pour 2 raisons essentielles. Tout d' abord, les schémas d' alternances sont clairement situés dans un cadre linguistique contemporain à

coloration générative. En dernier lieu, les études de M. Gross ont aussi porté sur les variations internes à chaque alternance (par exemple l' effacement d' un argument ou l' ajout d' un modifieur au sein d' une forme alternée). Nous nous contentons de considérer le passage d' une forme ' de base' à une forme alternée ' brute' , sans exclure des enchaînements ou des combinaisons d' alternances.

Pour analyser la pertinence sémantique de ces classes, nous les avons comparées à celles de même niveau de WordNet: le recouvrement n' est que de 46.36%, en prenant toujours les recouvrements maximaux en cas de choix multiples. Nous pouvons en conclure qu' une base purement syntaxique ne permet pas de construire des classes sémantiques ' exploitables' (comme le sont celles de WordNet), malgré les avantages que l' on peut y trouver. Il y a au moins deux raisons à cet état de fait: trop de verbes qui auraient du rentrer naturellement dans ces classes n' y sont pas en raison de leur comportement syntaxique, et d' autre part, pour certaines classes, les critères sémantiques en jeu sont très spécifiques et peu à même d' être utilisés dans les travaux en traitement automatique du langage.

Les résultats ci-dessus indiquent que les verbes étudiés ont des comportements syntaxiques très diversifiés, avec une moyenne d' environ 2.5 verbes par classe. Au niveau sémantique, on peut conclure de même, que les verbes ont des caractéristiques fines qui les différencient les uns des autres. Par ailleurs, le peu de recouvrement avec l' équivalent de WordNet que nous avons développé pour nos verbes français indique que les propriétés sémantiques qui président à la formation de classes sur une base syntaxique, et donc au fonctionnement syntaxique, sont de nature différente.

En effet, dans WordNet, les propriétés sémantiques utilisées ont un caractère général et se basent sur un découpage du monde en catégories fondamentales. Dans notre approche, les propriétés mises en jeu sont parfois très spécifiques et inattendues. Nous avons noté en particulier l' émergence de propriétés sémantiques très spécifiques. Bien qu' au tout premier niveau de la description, des concepts relativement spécifiques comme *'projection d'un matériau liquide ou considéré comme tel sur une surface'* ont été relevés. Un tel niveau de précision au niveau des restrictions de sélection a pu être pris en compte à chaque fois que des régularités ont été observées.

Par ailleurs, en remettant en cause l' analyse statistique simple employée, nous avons développé une analyse en composantes principales. Les résultats n' ont pas été meilleurs, en particulier à cause du fait que les critères identifiés comme composantes principales par l' algorithme de classification, caractérisés par des groupes d' alternances, avaient peu de signification linguistique.

Nous sommes toutefois parvenus à créer des classes `exploitables' en alliant les résultats précédents avec des critères sémantiques de base, en particulier des considérations thématiques et de restriction de sélection (Saint-Dizier 98). La pertinence psychologique de ces classes est évaluée dans (Dubois, Mari et Saint-Dizier 97).

Les classes réalisées comprennent une moyenne de 7 verbes, ce qui leur donne une granularité qui semble bien ajustée à notre propos. Par 'g ranularité bien adaptée' , nous voulons dire que les verbes d' une même classe ont une signification à peu près équivalente par rapport aux tâches que peut réaliser l' indexation de documents. Au niveau de leur représentation sémantique, ces verbes ont des représentations équivalentes, à des aspects adverbiaux près (la

manière, par exemple). Pour chaque verbe, nous trouvons une liste d' alternances (avec quelques exceptions), des restrictions de sélection, la mention de classes de prépositions si approprié, et une grille thématique.

Nos travaux sont en partie comparables à ceux de Beth Levin, bien que l' on doive souligner deux différences majeures entre les deux démarches. Tout d' abord, les classifications pour l' anglais entreprises par Beth Levin ont été réalisées en ne considérant que des sous-ensembles d' alternances qui semblaient permettre la construction de classes ayant une bonne homogénéité sémantique. Pour ces mêmes verbes, d' autres alternances possibles n' ont pas été considérées. De plus, Beth Levin a réalisé manuellement un traitement des exceptions, introduisant dans ces classes des verbes qui n' acceptaient pas nécessairement toutes les alternances sélectionnées initialement. En dépit de ce que nous percevons comme des inconvénients, le mérite principal du travail de Beth Levin, qui n' avait pas de visée technologique, a été de dégager avec une grande précision les facteurs sémantiques qui avaient une influence sur le comportement syntaxique des verbes.

2.2 Enrichissement des classes sémantiques de verbes

Ces classes obtenues à ce stade sont constituées à partir d' observations dans des corpus et de façon introspective. Afin de pouvoir les utiliser dans l' extraction de connaissances dans des textes, il est nécessaire d'y ajouter différents types d' informations.

Nous avons considéré différents types de textes: textes de rapports techniques (EDF), publications scientifiques du laboratoire, textes de revues générales, texte de cuisine. Il est clair qu' il y a des différences substantielles selon le type du texte, aussi bien dans les usages, que dans les restrictions de sélections, que même dans les types de métonymies et de métaphores rencontrées. Contrairement aux idées reçues, les textes techniques foisonnent de métaphores et de métonymies, mais celles-ci sont spécialisées (par exemple: structure pour responsable ou membre de la structure) et fortement récurrentes.

2.2.1 Enrichissement par des données statistiques sur les emplois

Un premier type d' information est lié aux réalisations en langue. A partir de nos observations, ainsi que de traitements statistiques sur des corpus marqués, avec une procédure d' apprentissage simple, nous avons enrichi chaque classe d' une liste d' usages. Ceci nous a permis de dégager des restrictions de sélection plus fines et de pondérer leurs usages par des fréquences d' utilisation. Ces observations comprennent non seulement des formes alternées identifiées, mais aussi une grande variété d' altérations du sens (par co-composition, par exemple) dont des métonymies et des métaphores. Nous avons complété notre étude par des éléments quantitatifs permettant une meilleure analyse en phase d' extraction de connaissances.

C' est ainsi que nous avons observé, dans notre approche du sens, que dans des textes à caractère technique:

(1) 24% des sujets et seulement 12 et 8% des objets 1 et 2 sont en situation métonymique et

(2) 8% des sujets et 24 et 15% des objets 1 et 2 sont en situation métaphorique. Au niveau des classes générales de WordNet, il y a de grandes disparités. C' est ainsi que les classes génériques ont les distributions globales suivantes (métaphores, métonymies) tous arguments confondus:

Principales Familles de verbes			
N°	*Nom de la classe*	*(1)*	*(2)*
1	Soins du corps	8	12
2	Verbes de changement	32	24
3	Verbes de communication	12	32
4	Verbes de compétition	17	21
5	Verbes de consommation	28	30
6	Verbes de contact	1	8
7	Verbes de cognition	15	10
8	Verbes de création et destruction	12	19
9	Verbes de mouvement	23	14
10	Verbes psychologiques	34	12
11	Verbes d'état (procédures, BE)	1	3
12	Verbes de perception	19	9
13	Verbes de possession	7	4
14	Verbes d'interactions sociales	33	29
15	Verbes liés à l'expression du temps	0	0
16	Verbes aspectuels et de l'action	1	20
17	Verbes qui expriment la causalité	17	12

Comme on peut le constater, les distributions sont très diversifiées. Cette étude a été réalisée sur 80 pages de textes variés. Nous disposons ainsi d'indications assez précises, qualitatives et quantitatives, sur les différentes réalisations syntaxiques directes et indirectes d'un sens de verbe, ou plus exactement, d'une classe sémantique de verbes.

Ces données permettent tout d'abord d'évaluer l'importance des phénomènes métonymiques et métaphoriques dans les textes. Ces considérations, dans une implémentation, permettent de mieux guider la stratégie d'analyse: selon le taux potentiel de métonymies ou de métaphores prévisible, une importance plus ou moins grande pourra être accordée à ce traitement coûteux en calcul. Nous considérons ces résultats comme des heuristiques élémentaires.

2.2.2 Enrichissement conceptuel des verbes

Un second type d'informations ajoutées à ces classes consiste en des représentations cognitives du sens, réalisées par le biais de formules sous-spécifiées (Saint-Dizier 99) de la Structure Lexicale Conceptuelle (LCS) (Jackendoff 90) (Dorr et Katsova 98). Le caractère sous-spécifié des formules en LCS permet d'introduire une dynamique avancée de la compositionalité, prenant en compte la diversité des réalisations linguistiques rencontrées dans les textes. Les liens entre WordNet et la LCS sont décrits dans (Dorr 99).

Prenons par exemple le verbe *remuer*. Dans notre approche, le sens primitif est celui d'un verbe de mouvement ('sur place' ou local). Sa représentation en LCS est la suivante:

λ I, J, [$_{event}$ CAUSE([$_{thing}$ I],
[$_{event}$ INCREASE$_{+char, +ident}$([$_{prop}$ MVMT-OF$_{+loc}$ ([$_{thing}$ J],
[$_{path}$ INSIDE$_{+loc}$([$_{place}$ LOC-OF([$_{thing}$ J])])])].

Cette représentation exprime simplement que I (l' agent) augmente le mouvement de J dans le lieu où J se trouve. On décrit ainsi un mouvement sur place, sans changement de lieu. Cette représentation prend en charge ce qui est de nature prédicative. Il est clair que d' autres éléments sémantiques (par exemple des paires attribut-valeur) sont nécessaires pour traiter de ce qui n' est pas prédicatif.

Ce verbe a des transpositions métaphoriques vers les domaines plus abstraits de la cognition et du psychologique: *remuer quelqu'un* signifie alors *secouer* ou *émouvoir* cette personne. On obtient ainsi en tout, au moins trois sens de remuer: un sens de base, et deux dérivés. Parmi les dérivés, nous présentons ici la représentation du sens dans le domaine psychologique. Cette représentation fait apparaître le changement de domaine ontologique, ainsi que la reformulation certaines fonctions dans le nouveau domaine. Pour le domaine psychologique, nous obtenons ainsi la représentation suivante:

λ I, J, [$_{event}$ CAUSE([$_{thing}$ I],
[$_{event}$ INCREASE$_{+char, +ident}$([$_{prop}$ PSY-AFFECT-OF$_{+psy}$ ([$_{thing}$ J],
[$_{path}$ BY-MEANS-OF$_{+psy}$([$_{place}$ QR(agentif, psy, emotion)])])].

Où la fonction QR(Role, Type, Mot) va chercher dans le rôle agentif de la structure Qualia (Pustejovsky 95) du mot Mot un prédicat de type +psy. C' est précisément dans cette dernière partie de la formule qu' est contenue l' information permettant de distinguer le sens en question comme se situant dans le domaine psychologique. De plus, cette représentation est partiellement sous-spécifiée au niveau de l' appel de la fonction QR: la valeur précise du champ dépendra de ce qui sera extrait du rôle agentif du mot *émotion*. La représentation ci-dessus se paraphrase en: I est la cause de l' accroissement de la ' pression' psychologique de J au moyen d' éléments susceptibles d' accroître son émotion (trouvés dans la Qualia de émotion).

A ce stade de l' étude, les trois options interprétatives que l' on peut donner de ces glissements métaphoriques nous paraissent équivalentes. Nous avons au moins 3 options: (1) considérer les 3 usages comme la spécification au sein d' une même classe d' un même sens sous-specifié, (2) insérer les 3 sens différents générés dans leurs classes respectives, (3) considérer qu' il s' agit d' un sens unique qui s' insère dans 3 classes différentes.

De plus, cela va nous permettre de faire 3 prédictions différentes et canaliser les développements ultérieurs:

(1) monotonicité: il est probable, comme maintes descriptions le montrent déjà, que des sous classes se présenteront comme un approfondissement d' un des composants sémantiques (ex. la manière) d' une classe plus abstraite,
(2) il est possible d' expliciter les liens entre les classes d' une même niveau d' abstraction et par là, entre les différents sens d' un même verbe susceptible d' apparaître dans différentes classes. Notons aussi l' analyse que l' on peut faire de composants LCS similaires entre calsses, qui peuvent introduire des relations ' transversales' ,
(3) on peut illustrer explicitement les mécanismes de co-composition par des représentations plus riches que de simples restrictions de sélection.

3. Extraction de connaissances dans des textes

Nos travaux ici illustrés, sont valorisés par le développement de méthodes linguistiques et informatiques d' extraction de connaissances dans des textes autour de la structure prédicat-argument (Pugeault et al. 1994). Nous procédons comme suit: une analyse ascendante (ou ' shallow') est réalisée d' un texte à traiter. Nous reconnaissons ainsi en totalité ou en partie des structures nominales ainsi que des verbes. Nous avons ensuite développé une sorte de système expert qui sait reconnaître des arguments d' un verbe, et tenté une composition sémantique de ces éléments, basés sur les données présentées ci-dessus.

Le formalisme général d' une forme prédicat-argument extraite est le suivant:
<Représentation du prédicat>(<role thématique ' argument' *>)

Par exemple, après composition on obtient pour la phrase extraite:
' Remuer le produit pendant 5 minutes'
Nous obtenons:

$[_{event}$ CAUSE($[_{thing}$ _],
$[_{event}$ INCREASE$_{+char, +ident}$($[_{prop}$ MVMT-OF$_{+loc}$ ($[_{thing}$ thème ' produit'],
$[_{path}$ INSIDE$_{+loc}$($[_{place}$ LOC-OF($[_{thing}$ thème ' produit'])])])].

4. Conclusion

Les données que nous avons présentées brièvement ici allient des considérations qualitatives et quantitatives qui s' appuient à la fois sur des descriptions linguistiques symboliques et des considérations statistiques. Ces dernières permettent de bien prendre en compte la grande diversité des réalisations linguistiques tout en canalisant ces réalisations vers une représentation uniformisée à caractère symbolique, nécessaire dans un système d' extraction de connaissances.

Les observations de corpus, permettent donc de mieux apprécier les variations d' usages dans des textes, et, de surcroît, par une légère généralisation sur celles-ci, de rendre compte à un niveau local et assez élémentaire de phénomènes tels que métaphores et métonymies. Ces généralisations sont essentiellement le passage d' un verbe à une classe de taille réduite qui l' englobe. Par exemple, si un verbe de mouvement accepte une certaine métaphore, l' idée est de déterminer la classe maximale (mais pas trop large) de verbes de mouvements qui va aussi accepter cette métaphore. Pour ce faire, on peut soit faire une analyse de corpus, ou bien avoir une approche introspective. On l' imagine, les exceptions sont nombreuses et la tâche délicate. Ces généralisations ont clairement besoin d' être validées et doivent donc être réalisées avec prudence, contrairement aux affirmations excessives de (Lakoff et al. 80) sur le caractère systématique des métaphores et des métonymies (par exemple contenant-contenu, concret-abstrait). Des approches locales, telles que celles de (Nunberg 95) et (Strigin 98), permettent de mieux maîtriser les alternances sémantiques observées.

Réferences

Dorr, B., Katsova, M., (1998), Lexical Selection for Cross-Language Applications: Combining
LCS with WordNet, 3rd conf. Machine Translation, Lahorne, PA.

Dorr, B., (1999), Large-scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation, Journal of Machine Translation, 12(1).

Dubois, D., Mari, A., Saint-Dizier, P., (1997), Quelques principes psycholinguistiques et formels pour la mise en oeuvre de la générativité en sémantique lexicale, AIDRI, Univ. de Genève.

Jackendoff, R., (1990), Semantic Structures, MIT Press.

Gayral, F., Saint-Dizier, P., Peut-on couper à la polysémie verbale ?, actes TALN 99, Cargèse.

Lakoff, G., Johnson, M., (1980), Metaphors we Live by, Chicago Univ. Press.

Levin, B., (1993), Verb Semantic Classes: a Preliminary Investigation, Chicago University Press.

Mari, A., Saint-Dizier, P., (1997), Générativité: au dela d' une théorie des types, TALN97, Grenoble.

Numberg, G., (1995), Transfer of Meaning, Journal of Semantics 12.

Pugeault, F., Saint-Dizier, P., Monteil, MG., (1994), Knowledge Extraction from Texts: a Method for Extracting Predicate-argument Structures, Coling, Kyoto.

Pustejovsky, J., (1995), The Generative Lexicon, MIT Press.

Saint-Dizier, P. (1998), Alternations and Verb Semantic Classes for French, in Predicative Forms for NL and LKB, P. Saint-Dizier (ed), Kluwer Academic.

Saint-Dizier, P., (1999), Underspecified Lexical Conceptual Structures for Sense Variations, Workshop on Lexical Semantics, Tilburg.

Strigin, A., (1998), Lexical Rules as Hypothesis Generators, Journal of Semantics 15.