Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ?

Vincent Claveau* — **Pascale Sébillot*** — **Pierrette Bouillon****
Cécile Fabre***

* *IRISA, Campus de Beaulieu, 35042 Rennes cedex, France*

{Vincent.Claveau,Pascale.Sebillot}@irisa.fr

** *TIM/ISSCO, Université de Genève, 40 Bvd du Pont-d'Arve, CH-1211 Genève 4, Suisse*

Pierrette.Bouillon@issco.unige.ch

*** *ERSS, Université de Toulouse II, 5 allées A. Machado, 31058 Toulouse cedex, France*

cfabre@univ-tlse2.fr

RÉSUMÉ. *Dans cet article, nous montrons, à travers l'exposé de quatre expériences d'apprentissage manipulant diverses informations catégorielles et sémantiques relatives aux mots d'un corpus, qu'il est possible d'acquérir automatiquement certains éléments du Lexique Génératif (sous forme de listes de couples nom-verbe dont les constituants sont liés par un des rôles de la structure des qualia), présentant un grand intérêt applicatif, en particulier en recherche d'information. Si l'exploitation des seules informations catégorielles par une méthode d'apprentissage de type programmation logique inductive offre déjà de bons résultats, nos travaux prouvent que le meilleur compromis qualité de l'apprentissage/coût de la méthode est obtenu en utilisant un étiquetage catégoriel couplé à un étiquetage sémantique des mots autres que les noms.*

ABSTRACT. *This paper demonstrates the feasibility of automatic acquisition of generative lexicons from corpora through the report of four experiments in machine learning in which various levels of word tagging (categorial and semantic) are handled. The lexical information that is learnt consists of lists of noun-verb couples related by one of the roles of the qualia structure. They provide linguistic knowledge useful in many applications such as information retrieval. We first show that satisfactory results are obtained on the basis of categorial information only, exploited by means of a learning method in the Inductive Logic Programming framework. We further demonstrate that the balance between quality and cost of the learning method is reached by the combination of a categorial tagging and a semantic tagging of words others than nouns.*

MOTS-CLÉS : *acquisition de lexiques sémantiques, lexique génératif, structure des qualia, apprentissage symbolique, programmation logique inductive.*

KEYWORDS: *semantic lexicon acquisition, generative lexicon, qualia structure, symbolic learning, inductive logic programming.*

1. Introduction

De nombreuses applications de traitement automatique des langues (TAL) nécessitent des informations lexicales sémantiques. Pour pouvoir disposer de telles données, deux voies sont possibles : soit utiliser des bases lexicales générales préexistantes, soit acquérir sur des corpus du domaine de l'application visée les ressources lexicales utiles.

Le thésaurus WordNet a, par exemple, servi de base à plusieurs travaux applicatifs [VOO 94, FEL 98, SME 99]. Cependant, certains auteurs réfutent la thèse de la pertinence de l'utilisation de telles ressources mutualisables, l'objection principale opposée à cette démarche étant qu'elle fait l'hypothèse qu'une ressource lexicale générale est valable hors contexte. En effet, de nombreuses études (cf. par exemple [BOU 97]) ont montré que la définition des relations de proximité sémantique ne peut pas être menée hors domaine mais doit au contraire s'appuyer sur les caractéristiques du corpus de travail. De plus, de façon plus globale, l'utilisation systématique de WordNet ou d'une autre base de ce genre en TAL est sujette à caution : dans quelle mesure un modèle sémantique conçu *a priori* s'avère-t-il adéquat pour représenter le fonctionnement de domaines particuliers ? Cette question de fond n'est pas toujours soulevée, et n'est en tous cas pas encore résolue, par ceux qui utilisent ce type de ressources.

Pour pallier ce problème, certains proposent de spécialiser des bases lexicales générales [BUI 97] grâce à des textes du domaine. Une autre solution consiste à acquérir, à partir de corpus, l'intégralité des connaissances lexicales sémantiques requises. De très nombreux travaux ont déjà été réalisés sur le sujet, essentiellement dans le cadre de l'apprentissage statistique (cf. [GRE 94b], par exemple, ou [HAB 97] et [PIC 97] pour des états de l'art du domaine). Ces recherches, généralement positionnées dans le cadre de la linguistique harrissienne [HAR 89], visent à extraire des informations syntagmatiques et paradigmatiques sur les unités lexicales, étudiant respectivement les mots qui apparaissent dans les mêmes fenêtres ou les mêmes contextes syntaxiques que l'unité considérée (affinités du premier ordre pour reprendre les termes de Grefenstette [GRE 94a]), ou les mots qui génèrent les mêmes contextes que le mot cible (affinités du second ordre). Par exemple, [BRI 97] et [FAU 99] tentent d'apprendre automatiquement des structures argumentales et des restrictions sélectionnelles ; [AGA 95] et [BOU 97] construisent des classes sémantiques ; [HEA 92] et [MOR 97] se focalisent sur un type particulier de relation lexicale, telle que l'hyperonymie. Certains de ces travaux ont également pour but d'obtenir automatiquement des représentations lexicales sémantiques plus complètes [GRE 94b, PIC 00].

Depuis plusieurs années, des recherches sur l'acquisition en corpus de lexiques sémantiques ont également vu le jour dans le cadre de l'apprentissage symbolique [WER 96]. C'est dans cet axe que se situent les travaux que nous exposons dans cet article. Notre objectif est en effet d'apprendre sur corpus un type de lexique sémantique particulier, le Lexique Génératif de Pustejovsky (LG ; [PUS 95, BOU 01a]), ou plus précisément certains éléments de ce lexique qui nous semblent particulièrement

pertinents pour la recherche d'information (RI)[1]. Dans LG, l'une des composantes, la structure des qualia, définit par des formules prédicatives essentiellement verbales les différentes facettes de la sémantique des noms. Le rôle télique indique le but ou la fonction du nom (*couper* pour *couteau* par exemple), l'agentif son mode de création (*construire* pour *maison*), le constitutif ses parties (*anse* pour *tasse*), et le rôle formel sa catégorie sémantique (*contenir (de l'information)* pour *livre*). Dans ce modèle, les noms (N) ne sont donc pas simplement liés via des relations lexicales traditionnelles (telles que la méronymie, l'hyperonymie, etc.) à d'autres noms, mais également à des verbes (V), nécessaires pour expliquer leurs comportements syntaxiques et sémantiques. Ce lien nomino-verbal est particulièrement intéressant dans le cadre de la RI et permet des reformulations d'index pertinentes et contrôlées, et actuellement peu usitées. Par exemple, le lien télique entre le N *jaugeur* et le V *mesurer* permet d'accéder à des extensions inter-catégorielles du type *jaugeur de carburant – mesurer du carburant.*

Cette hypothèse selon laquelle le lien sémantique nom-verbe peut être exploité en RI n'est certes pas nouvelle. Grefenstette, par exemple, montre l'importance de tels liens syntagmatiques pour aider à préciser et à désambiguïser les noms contenus dans des requêtes courtes [GRE 97]. Nous proposons, pour notre part, un moyen de systématiser une telle proposition et de définir les paires pertinentes : nous ne retenons, parmi les paires N-V possibles pour la reformulation sémantique, que celles décrites dans la structure des qualia de LG. Ce choix repose sur des considérations pratiques et théoriques. Sur le plan pratique, la validité de cette méthode a déjà été partiellement testée [FAB 96, FAB 99]. Sur le plan théorique, LG est une théorie des mots en contexte : elle définit des représentations lexicales sous-spécifiées qui acquièrent leurs spécifications sur corpus. Ces représentations peuvent ainsi être vues comme une façon de structurer l'information dans un corpus et c'est en ce sens que les relations définies dans LG sont exploitables en RI.

Étant donnée la pertinence pour la RI des liens N-V définis dans la structure des qualia de LG, il convient de mettre au point une méthode pour les acquérir en corpus, puisque de tels liens sont par nature définis en contexte. Les premières expériences menées dans cette direction, décrites dans [PUS 93], sont basées sur l'hypothèse que l'extraction de la structure des qualia d'un nom peut être réalisée en définissant un ensemble de structures syntaxiques liées aux rôles qualia. Nous proposons d'aller un pas plus loin car nous n'avons pas d'*a priori* concernant les structures susceptibles de porter ces rôles dans un corpus donné. Par conséquent, nous souhaitons développer une méthode d'apprentissage symbolique (de type programmation logique inductive [MUG 94]) qui produit automatiquement des règles générales capables d'expliquer ce qui, en termes de contexte environnant, distingue, dans la partie du corpus utilisée pour l'apprentissage, les exemples de paires N-V dont les constituants sont liés par un

1. Voir [BOU 00].

des rôles qualia des autres paires[2]. Toutefois, si les règles apprises sont explicatives du concept de rôle qualia, notre objectif n'est pas de tenter de formaliser la théorie du LG, c'est-à-dire, par exemple, de verbaliser la théorie à l'aide d'un nombre minimal de règles qui caractérisent les différents rôles qualia ; nous voulons plutôt obtenir un ensemble de règles, en nombre quelconque, les plus robustes possible, qui permettent de constituer des lexiques sémantiques. Plus précisément, les caractéristiques dont nous souhaitons doter notre méthode d'apprentissage sont les suivantes : d'une part, elle doit être fiable, c'est-à-dire qu'elle a pour but de fournir des règles générales qui, une fois appliquées sur le corpus, permettent d'obtenir des couples N-V effectivement liés par un des rôles de la structure des qualia. D'autre part, elle doit être réutilisable sur différents corpus dans des applications de TAL (recherche d'information ou autres) à moindre coût, ce qui implique que sa mise en œuvre soit suffisamment légère et n'impose pas un prétraitement trop lourd du texte.

Dans cet article, nous décrivons quatre expériences d'apprentissage qui utilisent différents types d'éléments de contexte pour décrire les paires N-V dont les constituants sont ou non reliés par un rôle qualia. Nous indiquons pour chacune d'elles les coûts et les résultats obtenus en termes d'acquisition automatique de lexiques sémantiques. Nous résumons d'abord successivement deux expériences préliminaires (voir [BOU 01b]), qui utilisent pour contexte des informations qui relèvent respectivement de l'étiquetage catégoriel du corpus et des étiquetages catégoriel et sémantique combinés. Nous montrons ensuite comment exploiter l'étiquette sémantique du N et du V et les problèmes d'efficacité qui en résultent au niveau de l'algorithme d'apprentissage. Enfin, nous étudions une alternative qui consiste à exclure volontairement l'information sémantique des noms pour simplifier l'apprentissage. Pour ces diverses tentatives, nous nous intéressons aussi à la validité linguistique des règles apprises. Nous concluons cet article en exposant les idées-clés que ces apprentissages successifs nous ont permis de formuler quant à l'apprentissage de LG sur corpus et en présentant nos perspectives de travail.

Pour alléger notre propos, nous appelons, dans la suite de l'article, *couples qualia* les couples N-V que l'on cherche à acquérir, c'est-à-dire ceux dont les constituants sont liés par l'un des rôles de la structure des qualia ; les autres couples sont dits *non qualia*. De même, lorsque nous parlons d'apprentissage de lexiques génératifs (ou sémantiques), nous faisons référence à l'acquisition de listes de couples N-V qualia[3].

2. Ce travail bénéficie du concours de l'Agence universitaire de la francophonie (AUF) (Action de recherche partagée *« Acquisition automatique d'éléments du Lexique Génératif pour améliorer les performances de systèmes de recherche d'information »*, réseau FRANCIL).

3. Nous cherchons donc seulement à obtenir des couples N-V dont les constituants sont liés par un des rôles de la structure des qualia, sans distinguer ces rôles ni tenter de construire entièrement la structure des qualia de ces N.

2. Acquisition de couples qualia à partir d'informations catégorielles

Cette première expérience, tout comme les trois autres décrites dans les sections suivantes, a donc pour but principal d'apprendre, à partir d'un corpus, des règles qui caractérisent les couples N-V qualia par rapport aux autres couples N-V ; elle consiste ensuite à appliquer ces règles sur le corpus afin d'obtenir des couples N-V qualia, qui pourraient servir à alimenter des lexiques sémantiques de type génératif. Tels quels, ces couples peuvent être, par exemple, utilisés en RI pour étendre des index et augmenter les performances des systèmes.

Avant de présenter l'expérience proprement dite et ses résultats, nous commençons cette section par l'exposé de notre cadre de travail, c'est-à-dire par la description de notre corpus d'étude et des bases de la programmation logique inductive (PLI ; [MUG 94]), cadre d'apprentissage symbolique dans lequel se situe la méthode que nous avons développée. Nous décrivons ensuite la mise en place de cette première méthode d'apprentissage fondée sur l'utilisation de l'étiquetage catégoriel des mots du corpus et, plus particulièrement, sur celui des mots environnant les N et les V formant ou non des couples qualia dans la partie du corpus qui sert à l'apprentissage. Nous n'abordons pas ici les aspects purement techniques de cette méthode (cf. sur ce point [BOU 00, SÉB 00]) et nous nous focalisons davantage sur les résultats en termes de qualité de l'apprentissage des couples qualia à l'aide des règles générales inférées, ainsi que sur la pertinence linguistique de ces règles.

2.1. *Le corpus et le cadre d'apprentissage*

Le corpus que nous utilisons est un manuel de maintenance d'hélicoptères en français qui nous a été fourni par MATRA CCR Aérospatiale. Il contient plus de 104 000 occurrences de mots, soit une taille d'environ 700 ko. Ce corpus a plusieurs caractéristiques intéressantes pour la tâche que nous visons : il est très homogène dans ses structures syntaxiques et son vocabulaire ; il compte un grand nombre de termes très concrets (*vis, porte, etc.*) qui sont fréquemment utilisés au sein d'une phrase avec des verbes marquant leur rôle télique (*serrer les vis*, etc.) ou leur rôle agentif (*effectuer un réglage*, etc.). Le corpus MATRA CCR semble donc parfaitement se prêter à l'acquisition de couples N-V par une technique d'apprentissage symbolique basée sur des exemples, comme nous nous proposons de le faire. Le but de l'apprentissage automatique est en effet de construire automatiquement des programmes à partir d'exemples, que l'on sait positifs ou négatifs, de leur fonctionnement [MIT 97]. Parmi les différentes techniques existant dans ce domaine, nous avons choisi la *programmation logique inductive* (PLI) [MUG 94] pour acquérir à partir du corpus des couples N-V tels que le V appartienne à la structure des qualia du N.

La PLI a pour but d'induire des théories – exprimées par des programmes logiques sous forme de clauses de Horn – à partir d'exemples et d'un ensemble de connaissances préalables (*background knowledge*). Plus précisément, un algorithme de PLI essaie de construire, à partir des connaissances préalables, des hypothèses génériques

qui expliquent les exemples positifs (E^+) tout en rejetant les exemples négatifs (E^-) (du moins le maximum d'exemples négatifs, un peu de bruit pouvant être toléré). Ces hypothèses sont le plus souvent générées en s'appuyant sur un langage (sorte de biais syntaxique), donné par l'utilisateur et qui assure ainsi la production d'hypothèses bien formatées par rapport au problème posé. C'est le caractère explicatif de la PLI qui a guidé notre choix : elle ne se contente pas, contrairement à d'autres méthodes de type boîte noire, de construire un prédicteur, mais elle fournit aussi une théorie fondée sur les données.

Dans notre cas, les exemples positifs sont des paires N-V dont les constituants sont en relation syntaxique au sein d'une même phrase et sont liés par l'un des rôles qualia (par exemple, *frein lâcher* dans *lâcher le frein de parking*). De même, nos exemples négatifs sont des couples N-V qui cooccurrent dans des phrases du corpus MATRA CCR, mais sans que leurs éléments ne soient reliés par un tel rôle, par exemple *frein relâcher* dans *relâcher la commande de frein de parking*. Les hypothèses que nous cherchons à inférer à partir de ces E^+ et E^- sont des règles ou clauses, obtenues à l'aide d'un algorithme de PLI par généralisation contrôlée des E^+, et qui expliquent ce qui caractérise les couples qualia par rapport aux autres. Ces règles, une fois appliquées sur le corpus, permettent ainsi d'extraire d'autres paires liées de la même manière.

Comme notre objectif est de développer une méthodologie capable de construire de façon peu coûteuse mais fiable des lexiques génératifs, notre première expérience de mise au point d'une méthode d'apprentissage automatique de type PLI se base sur l'utilisation du seul étiquetage catégoriel du corpus.

2.2. *Première expérience d'apprentissage*

Pour cette première expérience, nous avons donc étiqueté catégoriellement le corpus MATRA CCR Aérospatiale à l'aide d'outils développés dans le cadre du projet MULTEXT [ARM 96]. Ces outils permettent tout d'abord de segmenter le texte en unités lexicales ; les unités sont ensuite analysées et lemmatisées, puis finalement désambiguïsées grâce à l'outil *Tatoo*, un étiqueteur à chaînes de Markov cachées [ARM 95]. Ce traitement permet ainsi d'associer aux mots du corpus des informations – exploitées par la suite par notre système d'apprentissage – sur la catégorie morpho-syntaxique (nature, genre, nombre, etc.) de chaque mot, et ce, avec une grande précision, puisque sur un échantillon-test de texte de 4 000 mots, seuls 2 % ont été incorrectement étiquetés.

La première étape de la mise au point de notre méthode d'apprentissage des couples qualia consiste à constituer, à partir du corpus, un ensemble d'E^+ de couples N-V et un ensemble d'E^-. Dans ce but, nous commençons par associer aux mots les plus courants du corpus les 10 verbes qui leur sont le plus fortement liés, en termes de χ^2, au sein des phrases de ce corpus. Nous obtenons ainsi des couples effectivement liés par une relation de type qualia (*écrou, serrer*), mais aussi des paires clairement

sans rapport sémantique (*roue, prescrire*). Pour chacun des couples potentiellement qualia, un expert examine alors toutes les phrases dans lesquelles il apparaît pour indiquer si, pour cette occurrence précise, le couple dénote bien une relation qualia [4]. Si c'est le cas, l'occurrence va servir à former un exemple positif pour l'apprentissage, sinon un exemple négatif (les occurrences des couples non potentiellement qualia étant aussi transformées en exemples négatifs). Pour ce faire, la phrase est alors transformée automatiquement, via un programme *Perl*, en un prédicat (précédé de :- si c'est un exemple négatif) du type :

positif(tag[5]_du_mot_avant_N, tag_du_mot_après_N, tag_du_mot_avant_V, tag_de_V, distance, position).

où distance indique le nombre de verbes conjugués entre N et V dans la phrase et position spécifie si V apparaît avant N (codé pos[6]) dans la phrase ou le contraire (codé neg). Par exemple, à partir de la phrase « *L'installation se compose : de deux atterrisseurs* **protégés** *par des* **carénages**, *fixés et articulés…* » et du couple qualia **carénage-protéger** qui y apparaît en gras, est formé l'E^+ :

positif(p_par, virg, nc_pl, verb_adj, 0, pos).

où p_par indique que le mot précédant le N est la préposition *par* (les déterminants et certaines autres catégories étant ignorés), virg qu'une virgule suit le N, nc_pl qu'un nom commun au pluriel précède le V, vrb_adj que le V est au participe passé, 0 et pos dénotant que le V précède le N sans que ces éléments ne soient séparés par un verbe.

Le choix de la forme des clauses-exemples s'appuie sur l'intuition que le contexte proche du N et du V (les trois premiers champs des clauses), les informations sur le N ou le V (quatrième champ) et les positions relatives du N et du V dans la phrase (deux derniers champs), sont de bons indicateurs de l'appartenance du V à la structure des qualia du N. Cependant la pertinence des champs effectivement retenus a été mesurée grâce à un programme *Perl* qui prend successivement plusieurs éléments de contexte parmi un ensemble de propositions et qui teste la qualité de l'apprentissage réalisé à l'aide de cette sélection (cf. [BOU 00, SÉB 00] pour une description précise de la façon dont la qualité de l'apprentissage peut être évaluée, explication que nous nous contentons d'aborder rapidement ci-dessous).

Plus de 4 000 clauses (exemples) positives et environ 7 000 négatives ainsi formées ont alors été fournies en entrée de l'algorithme Progol [MUG 95], pour qu'il effectue les généralisations possibles des E^+ et produise des règles générales expliquant le concept de couple qualia. Progol est une implémentation de l'algorithme de PLI proposé par Muggleton qui a déjà servi dans de nombreux domaines et donne des résultats comparables aux autres algorithmes de PLI [ROB 98]. Nous obtenons ainsi des clauses généralisées (66 au total) de la forme :

positif(A,B,C,D,E,F) :- aux_etre(C), prepositionpar(A), pres(E).

4. Les phrases sont présentées automatiquement à l'expert, mais la validation est manuelle.
5. Étiquette.
6. Pour positif, c'est-à-dire respectant l'ordre VN.

qui signifie que les paires N-V (i) dont le mot précédant N est la préposition *par* (prepositionpar(A)), (ii) dont le mot précédant V est un auxiliaire (aux_etre(C)), (iii) et dont les constituants N et V ne sont pas séparés par un autre verbe (pres(E)), sont de type qualia.

La qualité intrinsèque de l'apprentissage ainsi produit est mesurée en combinant le pourcentage d'exemples positifs et d'exemples négatifs couverts par les clauses généralisées en un coefficient de Pearson[7]. Ce coefficient, s'il est proche de 1, indique un apprentissage parfait ; dans notre cas, 88 % des exemples positifs et seulement 5 % des exemples négatifs sont couverts, ce qui correspond à une valeur du coefficient de Pearson de 0,84.

On mesure aussi la qualité de l'apprentissage pour la tâche qui nous intéresse, à savoir la capacité des clauses générales obtenues à constituer un lexique de couples N-V qualia. Cette validation empirique de l'apprentissage est réalisée en appliquant les clauses généralisées sur une partie du corpus, et en s'intéressant aux couples qualia qu'elles permettent de détecter. Pour des raisons de faisabilité, nous ne considérons, pour cette validation, que sept noms jugés significatifs dans le domaine (*vis, écrou, porte, voyant, prise, capot, bouchon*). D'une part, un expert examine manuellement la partie du corpus et classe tous les couples réalisables à partir de ces sept noms en qualia ou non. Parallèlement, nous appliquons les règles générales sur la même partie du corpus et, pour chacun de ces noms, nous obtenons un certain nombre de verbes, détectés une ou plusieurs fois chacun (soit dans des phrases différentes, soit par des clauses différentes) comme jouant un de ses rôles qualia ; il est en effet possible de décider qu'un couple N-V est effectivement considéré comme qualia s'il est détecté au moins un certain nombre de fois. Les résultats sont ensuite comparés à ceux de l'expert et synthétisés dans un coefficient de Pearson. En considérant qu'un couple doit être détecté au moins une fois pour être considéré comme qualia, nous obtenons 49 paires classées qualia à juste titre, 54 incorrectement détectées et 10 non trouvées, soit une valeur du coefficient de Pearson de 0,5138 ; si on considère qu'il faut 6 détections, on obtient respectivement 23, 4 et 36, soit une valeur du coefficient de 0,5209. À titre de comparaison, sur ces mêmes noms, la méthode statistique du χ^2, initialement utilisée dans la construction des exemples, donne respectivement 38, 124 et 21, soit une valeur du coefficient de Pearson de 0,1206.

Notre méthode d'acquisition, en se basant uniquement sur des informations catégorielles issues de l'étiquetage, mène donc à des résultats très encourageants : 83,05 % des paires qualia sont correctement détectées contre 64 % pour la méthode du χ^2. De plus, contrairement aux méthodes statistiques, elle met en exergue le caractère explicatif de la PLI. En effet, ce ne sont pas seulement des paires corrélées qui sont automatiquement extraites, mais aussi des règles permettant de distinguer une paire qualia d'une autre non qualia. Celles-ci ont d'ailleurs pour la plupart une réelle signification linguistique, même si elles exploitent surtout des marques superficielles d'ordre

7. Pearson = $\frac{(TP*TN)-(FP*FN)}{\sqrt{PrP*PrN*AP*AN}}$, où A = *actual* (réel), Pr = *predicated* (prédit), P = *positive*, N = *negative*, T = *true* et F = *false*.

des mots, les signes de ponctuation ou des chaînes de catégories grammaticales. En particulier, l'analyse des clauses révèle que :

– comme on pouvait s'en douter, la proximité est un critère important. 14 clauses parmi les 66 apprises expriment le fait que les N et V liés doivent être séparés par un seul élément au plus (soit une préposition, soit un nom propre, soit une virgule, etc.), par exemple : « *tâches* **à** *effectuer* »; « *écrous***,** *serrer* »; « *voyants* **joker** *allumés* ». 35 clauses contiennent l'information qu'il ne doit y avoir aucun verbe entre les deux constituants du couple ;

– 20 clauses montrent l'importance des marques de ponctuation. Dans notre corpus, elles sont en effet suffisantes pour identifier certains couples ; c'est le cas par exemple si le N et le V sont séparés par deux points, et si le N ou le V sont directement précédés ou suivis par une virgule : « ...**,** *écrou* **:** *serrer* au couple**,** » ;

– la position du couple N-V dans la phrase est souvent explicitée : 16 clauses généralisent le fait qu'un verbe à l'infinitif en début de phrase est un candidat sérieux. La plupart expriment dans une tournure impersonnelle un ordre référant à une action typique à faire sur un objet, par exemple : « *procéder* à un *essai* » ; « *éliminer* toute trace de *calcaire* ». C'est assez typique de notre corpus qui contient de nombreuses instructions d'utilisation ;

– enfin, les clauses montrent la pertinence de certaines constructions syntaxiques. L'une d'entre elles caractérise la tournure passive : N et V sont en relation si V est précédé de l'auxiliaire *être* et suivi de la préposition *par*. *A priori* plus surprenant, deux clauses spécifient que le N et le V qui suivent une conjonction de subordination sont pertinents. Cela généralise le fait que beaucoup de verbes, dans notre corpus, requièrent des complétives indiquant une action typique, comme « s'assurer que » ou « vérifier que » (« s'assurer **que** l'*alimentation* est *coupée* » ; « vérifier **que** le *feu* anti-collision *clignote* »).

Le fait que quelques exemples négatifs soient recouverts par les clauses généralisées indique cependant qu'il manque quelque chose dans nos exemples pour permettre de complètement définir le concept de paire qualia. Des informations syntaxiques ou sémantiques permettraient peut-être de mieux caractériser ce concept. Cela peut facilement s'illustrer par l'exemple suivant : dans les structures du type « verb-inf det N1 prep N2 », le verbe à l'infinitif et le nom N2 ne sont parfois pas en relation (*vérifier l'absence de corrosion*) et le sont à d'autres occasions, notamment quand N1 indique un groupe (*vider les deux groupes de réservoirs*). Un simple étiquetage catégoriel du corpus ne permet pas de rendre compte de cette différence. Nous l'avons donc étiqueté sémantiquement et avons conduit une expérience similaire en tenant compte de ces nouvelles informations. Nous décrivons dans la section suivante cette étape d'étiquetage sémantique et les résultats obtenus par notre méthode d'acquisition de couples.

3. Acquisition de couples qualia à partir d'informations catégorielles et sémantiques sur leur contexte d'apparition

Dans cette partie, nous décrivons rapidement la façon dont nous avons réalisé l'étiquetage sémantique du corpus MATRA CCR ; le lecteur pourra se reporter à [BOU 01b] pour plus de détails. Nous présentons ensuite l'apprentissage mené sur le corpus étiqueté catégoriellement et sémantiquement, les résultats obtenus et leur comparaison avec ceux de l'expérience précédente.

3.1. *Annotation sémantique du corpus*

La méthode d'annotation sémantique que nous utilisons repose sur trois hypothèses majeures : (i) les informations catégorielles peuvent aider à distinguer les sens des mots polyfonctionnels[8] [WIL 96, YAR 92, CEU 96], (ii) l'étiquetage catégoriel (non ambigu) peut être réalisé par un étiqueteur probabiliste (voir 2.2), et, ce qui est plus surprenant, (iii) les ambiguïtés sémantiques restantes peuvent aussi être résolues par un étiqueteur probabiliste.

L'étiquetage sémantique du texte nécessite dans un premier temps de construire manuellement, pour chaque catégorie de mot, un lexique contenant pour chaque entrée les différentes étiquettes qu'elle peut porter au sein du corpus. Cela implique de choisir pour chaque catégorie un jeu d'étiquettes sémantiques adapté.

Pour classifier les noms du corpus de manière systématique, nous avons utilisé, comme point de départ, les classes les plus génériques de WordNet [FEL 98]. Cependant, certaines de ces classes, inusitées dans notre corpus, ont été supprimées, alors que d'autres, très présentes, ont été raffinées en sous-classes plus précises (c'est le cas en particulier de la classe des objets concrets). Nous avons ainsi obtenu 33 classes, organisées en une hiérarchie représentée en figure 1 (les classes initiales de WordNet non usitées sont en italiques, et les étiquettes sémantiques choisies apparaissent entre parenthèses). Le tableau 1 donne, quant à lui, pour chaque classe, son effectif et quelques représentants. Environ 8,7 % des entrées du lexique des noms constitué sont ambiguës. Ces ambiguïtés correspondent le plus souvent à des phénomènes de polysémie complémentaire (par exemple, *enfoncement* peut indiquer à la fois un processus et son résultat ; il est donc classifié en **pro** et **sta**).

En ce qui concerne les verbes, la classification de WordNet a été jugée inadaptée du fait d'un trop grand éparpillement des classes. Nous avons donc adopté une partition minimale en sept classes dans laquelle très peu de verbes sont ambigus (seulement 6 sur près de 570). Les autres catégories de mots du corpus (prépositions, pronoms, etc.) ont aussi été organisées en classes et rangées dans un lexique ; là encore, comparativement aux noms, peu d'entrées pour ces catégories sont ambiguës.

8. C'est-à-dire qui ont plusieurs catégories, comme *règle* qui peut indiquer soit un nom, soit un verbe à l'indicatif.

Code	Définition	Effectif	Exemples
act	activité, nom d'action	13	description, maintenance
acy	activité humaine	15	réparation, visite
agt	cause, agent	5	contrainte, risque
art	artefact	358	filtre, piège
atr	attribut	17	aspect, interchangeabilité
chm	composé chimique	14	acétone, azote
cnt	conteneur	21	bol, bocal
com	communication	54	alarme, code
ent	entité	4	champignon, oiseau
frm	forme	35	arrondi, boursouflure
grp	groupe	18	alignement, gamme
grs	groupe social	3	personnel, équipage
hap	événement	64	anomalie, avarie
hum	humain	9	copilote, motoriste
ins	instrument	266	anémomètre, peson
loc	localisation	27	bord, dehors
mea	mesure, quantité	31	minimum, surplus
phm	phénomène naturel	31	chaleur, givre
pho	objet physique	24	corps, dépôt
pnt	point de localisation	19	angle, aplomb
por	partie, portion	26	branche, bras
pos	position d'un objet physique	14	emplacement, endroit
pro	processus	274	accumulation, décapage
prt	partie du corps	2	genou, main
psy	trait ou activité psychologique	30	besoin, connaissance
pty	propriété	52	irrégularité, longueur
qud	quantité définie	10	coefficient, double
rel	relation entre objet	34	altitude, dépassement
sta	état	45	ambiance, besoin
stu	matériau, matière	21	alliage, amiante
sub	substance	37	carburant, colle
tme	indication de temps	18	heure, an
unt	unité	24	ampère, bar

Tableau 1. *Définition, effectif et exemples des classes sémantiques des noms*

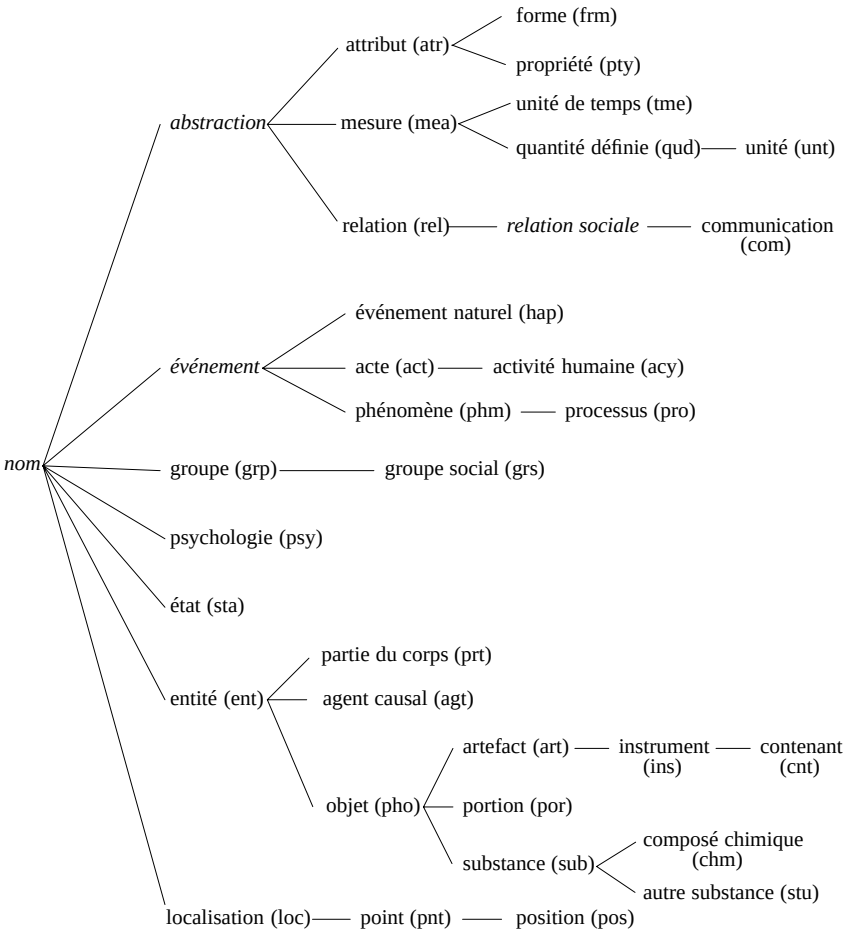

Figure 1. *Hiérarchie de classes pour l'étiquetage sémantique des noms*

L'étiquetage sémantique consiste alors à projeter sur chaque mot du corpus (déjà étiqueté catégoriellement) le contenu de l'entrée correspondante du lexique dont nous venons de décrire le mode de constitution. Les ambiguïtés sont ensuite résolues en utilisant, comme pour l'étiquetage catégoriel, un *tagger* à chaînes de Markov cachées. Comme mentionné ci-dessus, les ambiguïtés à résoudre sont principalement des problèmes de polysémie complémentaire, puisque les mots ont déjà subi une désambiguïsation catégorielle qui limite la polysémie contrastive (comme l'homographie). Une portion du texte de près de 6 000 mots a servi à mesurer la précision de notre étiquetage sémantique. Dans cet extrait, 7,78 % des mots étaient initialement ambigus, et l'étiquetage a permis de résoudre correctement 85 % de ces ambiguïtés.

3.2. *Seconde expérience d'apprentissage*

De la même manière que précédemment, les phrases comprenant les couples N-V repérés par le χ^2 (qui contiennent désormais des informations catégorielles et sémantiques) sont transformées en clauses-exemples. Dans le but de pouvoir comparer les résultats aux précédents, nous avons choisi d'adopter pour ces clauses une forme très similaire à celle utilisée auparavant (cf. page 735), soit :

positif(tag_sémantique_précédant_N, tag_sémantique_suivant_N,
tag_sémantique_précédant_V, tag_catégoriel_de_V, distance).,

où distance indique le nombre de verbes entre N et V, et, suivant son signe, indique l'ordre d'apparition de N et V dans la phrase (cela équivaut donc aux arguments distance et position des exemples de la première expérience). La phrase « *L'installation se compose : de deux atterrisseurs* ***protégés*** *par des* ***carénages****, fixés et articulés…* » est donc cette fois transformée en :

positif(preposition_maniere, ponctuation, artefact, verb_adj, 1 [9]).

C'est donc volontairement que nous n'exploitons pas ici toutes les informations que nous avons maintenant sur le corpus, notamment les étiquettes sémantiques de N et V ; notre but est, en effet, d'évaluer à travers nos diverses expériences d'apprentissage l'influence exacte de chaque type d'information.

Les 4 000 E^+ et 7 000 E^-, avec cette nouvelle forme, sont fournis à Progol pour obtenir des généralisations s'appuyant donc non plus seulement sur des informations catégorielles mais aussi sur les informations issues de l'étiquetage sémantique.

L'évaluation théorique de cette nouvelle phase d'apprentissage est conduite de la même manière que précédemment et les résultats se révèlent, comme attendu, meilleurs que les précédents : 346 clauses généralisées sont obtenues ; elles couvrent (i.e. expliquent) 89,9 % des exemples positifs et seulement 0,7 % des négatifs. Cela donne donc un coefficient de Pearson de 0,91, supérieur à celui de notre première expérience, et marque donc un apprentissage plus précis.

Voici par exemple une des clauses inférées par Progol : positif(A,B,C,D,E) :- vide(A), etat_verbe(C), modalite(B) ; elle signifie qu'est considérée comme qualia toute paire N-V dont (i) le N est en début de phrase (vide(A), rien ne précède le N), (ii) le mot suivant le N est un verbe modal (modalite(B)), et (iii) le mot précédant le V est un verbe d'état (etat_verbe(C)), aucune contrainte n'étant indiquée sur la distance ou l'ordre de N et V, ni sur la catégorie du V. Cette règle, par exemple, reconnaît les paires (*platine, déposer*) et (*train, sortir*) comme qualia dans les deux phrases suivantes extraites du corpus MATRA CCR : *« Les* ***platines*** *doivent être* ***déposées*** *s'il y a échange du réservoir. »* et *« Le* ***train*** *peut être* ***sorti*** *à l'aide de l'électropompe secours en cas de perte soit : de pression des pompes, de l'alimentation 28 V du circuit de commande. »*.

9. Pour pouvoir distinguer une distance nulle positive d'une distance négative, toutes les distances sont décalées d'une unité.

Le processus de validation empirique décrit en 2.2 est également utilisé ici pour évaluer notre technique d'acquisition automatique de couples qualia. Nous appliquons donc les 346 clauses sur la même partie du corpus et examinons les couples obtenus qui contiennent l'un des sept noms choisis. Comme précédemment, le nombre de détections de ces couples permet de choisir le meilleur compromis entre la quantité de couples retrouvés (rappel) et la qualité de ces couples (précision). Si un couple est considéré comme qualia dès qu'il est détecté au moins une fois, 57 paires sont correctement trouvées, 61 incorrectement et 13 ne sont pas détectées ; cela donne une valeur du coefficient de Pearson de 0,4678. Si le nombre de détections minimales pour considérer un couple comme qualia est fixé à 2, les résultats sont respectivement 47, 21 et 13, soit un coefficient de Pearson de 0,5817. Ces nombres permettent donc de constater que la valeur maximale du coefficient de Pearson est supérieure de 11,5 % par rapport aux résultats obtenus en n'utilisant que des informations catégorielles. De plus, le nombre de détections nécessaires pour obtenir ce maximum est inférieur à celui de l'expérience précédente (2 contre 6). Cela signifie que chacune des 346 clauses est une meilleure caractérisation, plus précise et plus fidèle, du concept d'appartenance d'un verbe à la structure des qualia d'un nom.

Comme pour la première expérience, une partie de ces clauses reçoit une interprétation linguistique pertinente. Bien sûr, nombre d'entre elles reprennent des schémas déjà explicités en 2.2, mais quelques-unes spécifient des propriétés sémantiques intéressantes sur les mots du contexte du couple N-V. Ainsi, les deux règles suivantes utilisent notamment les informations sémantiques des verbes et des prépositions précédant ou suivant les constituants du couple N-V :

– les verbes modaux comme *permettre*, *devoir* ou *pouvoir* sont de bons indicateurs de couples pertinents : « le *tableau* **doit** être *éclairé* », « l'*adhésion* **peut** être *atteinte* », etc. ;

– le type sémantique de la préposition peut aider à identifier des couples qualia, spécialement si la préposition indique la manière ou le but : « *fixer* **avec** leurs *vis* sans serrer », « **pour** *emmancher* l'arbre d'entraînement **dans** la *prise* de mouvement ».

L'étiquetage sémantique, même utilisé de manière partielle, apporte donc de meilleurs résultats, à la fois en termes de pertinence linguistique des règles apprises et, pour ce qui concerne plus spécifiquement notre but de constitution de lexiques génératifs, en termes de rappel et de précision pour l'acquisition des couples qualia à partir du corpus. Il requiert cependant un travail de classification coûteux, en particulier en ce qui concerne les noms. Nous présentons dans la partie suivante une troisième expérience utilisant les informations (catégorielles et sémantiques) disponibles non plus de manière partielle comme jusqu'ici mais de manière exhaustive.

4. Acquisition de couples qualia à partir d'informations catégorielles et sémantiques

L'apport de l'étiquetage sémantique étant démontré par les résultats de l'expérience précédente, nous essayons maintenant d'exploiter pleinement les informations sémantiques du corpus. Nous prenons donc en compte dans cette nouvelle expérience non seulement les étiquettes sémantiques et catégorielles des mots du contexte environnant du couple N-V, mais aussi celles du N et du V.

Pour des raisons de souplesse d'utilisation et de lisibilité, nos exemples sont toutefois codés d'une manière quelque peu différente de celle utilisée jusqu'à maintenant. Ainsi, un exemple positif tiré de la phrase « *L'installation se compose : de deux atterrisseurs* ***protégés*** *par des* ***carénages****, fixés et articulés…* » est, par exemple, simplement noté :
positif(n609,v609).,
et les informations concernant ce couple N-V et son contexte sont données par ailleurs, sous la forme :

```
tags(m609_8,tc_prep,ts_rman).
pred(n609,m609_8).
tags(m609_10,tc_wpunct,ts_virg).
suc(n609,m609_10).
tags(m609_6,tc_noun_pl,ts_art).
pred(v609,m609_6).
suc(v609,m609_8).
tags(n609,tc_noun_pl,ts_art).
tags(v609,tc_verb_adj,ts_acp).
distances(n609,v609,2,1).
```

tags donne des informations sur la catégorie et la classe sémantique d'un mot, pred et suc indiquent les prédécesseur et successeur d'un mot, et distances précise le positionnement et la distance entre deux mots. Les prédicats ci-dessus signifient que le mot m609_8 suit (suc(v609,m609_8)) le verbe V (ici v609) et précède (pred(n609,m609_8)) le nom N (n609) dans la phrase, et que ce mot est une préposition (tc_prep) de manière (ts_rman). Il y est indiqué de la même manière que V est un verbe dénotant une action physique (ts_acp) au participe passé (tc_verb_adj). Enfin, distances(n609,v609,2,1) indique la distance[10] en nombre de mots entre N et V, et la distance en nombre de verbes conjugués.

Par rapport à l'expérience précédente, le programme d'inférence manipule donc beaucoup plus d'informations puisqu'il tient compte désormais des étiquettes catégorielles et des étiquettes sémantiques de N et de V ainsi que de tous les mots qui leur sont contigus. De plus, il est libre de trouver des généralisations « contraignant » les valeurs d'un nombre quelconque de ces éléments et non, comme dans les deux expériences précédentes, d'un nombre quelconque des éléments choisis comme champs

10. Positive si V précède N, négative sinon, et décalée d'une unité pour différencier l'ordre dans le cas d'une distance nulle, comme en section 3.

fixes des clauses-exemples (6 dans la première et 5 dans la deuxième). D'importants problèmes d'efficacité et de rapidité en découlent. Ils ont été résolus en deux temps : tout d'abord par l'adoption d'Aleph[11], une autre implémentation de l'algorithme de Muggleton, plus rapide et plus souple que Progol utilisé auparavant, et ensuite par l'écriture d'un opérateur de raffinement bien adapté à notre tâche, dont nous expliquons le fonctionnement et l'intérêt ci-dessous.

4.1. *Efficacité de l'algorithme de PLI - opérateur de raffinement*

L'opérateur de raffinement est en fait le cœur de tout système de PLI puisque son rôle consiste à générer des hypothèses bien formées (i.e. respectant le biais syntaxique donné par l'utilisateur) et qui couvrent le maximum d'exemples positifs et le minimum d'exemples négatifs.

Pour expliquer son fonctionnement, décrivons plus précisément l'algorithme utilisé par Aleph (et de nombreux autres algorithmes de PLI), qui peut se découper en quatre étapes :

1) sélectionner un E^+ ;

2) construire la clause la plus spécifique à partir de cet exemple. Cette clause (appelée *bottom clause*) est en fait l'hypothèse la moins générale qui couvre l'exemple sélectionné. Le processus de construction de la *bottom clause* (appelé parfois saturation) est détaillé dans [MUG 95] ;

3) rechercher la meilleure hypothèse. Cette hypothèse doit être plus générale que la *bottom clause* ;

4) ôter les exemples couverts par l'hypothèse retenue et retourner au point 1.

L'opérateur de raffinement intervient lors de la troisième étape. Il permet de rechercher dans l'ensemble des hypothèses possibles la clause qui, relativement à une certaine fonction de score, se révèle la meilleure. Cet ensemble de clauses est un treillis ordonné par une notion de généralité. Au sommet de ce treillis se trouve la clause la plus générale (dans notre cas, c'est positif(A,B) : tout couple N-V est qualia) et la *bottom clause* est quant à elle la borne inférieure. La recherche de l'hypothèse maximisant la fonction de score se fait en parcourant le treillis entre ces deux clauses (cf. figure 2).

Le parcours se fait généralement de haut en bas, c'est-à-dire du plus général au plus spécifique. Ainsi, à partir d'une hypothèse se révélant trop générale (couvrant trop d'exemples négatifs) est générée[12] une autre hypothèse plus spécifique. Le score de cette dernière est alors évalué et le processus est réitéré jusqu'à être certain qu'il n'y ait pas dans le treillis une meilleure hypothèse. Il n'est généralement pas nécessaire de

11. Aleph, anciennement appelé P-Progol, est disponible sur le site http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/

12. C'est ce processus qui a conduit Shapiro à appeler ce système de génération d'hypothèses *opérateur de raffinement* [SHA 81].

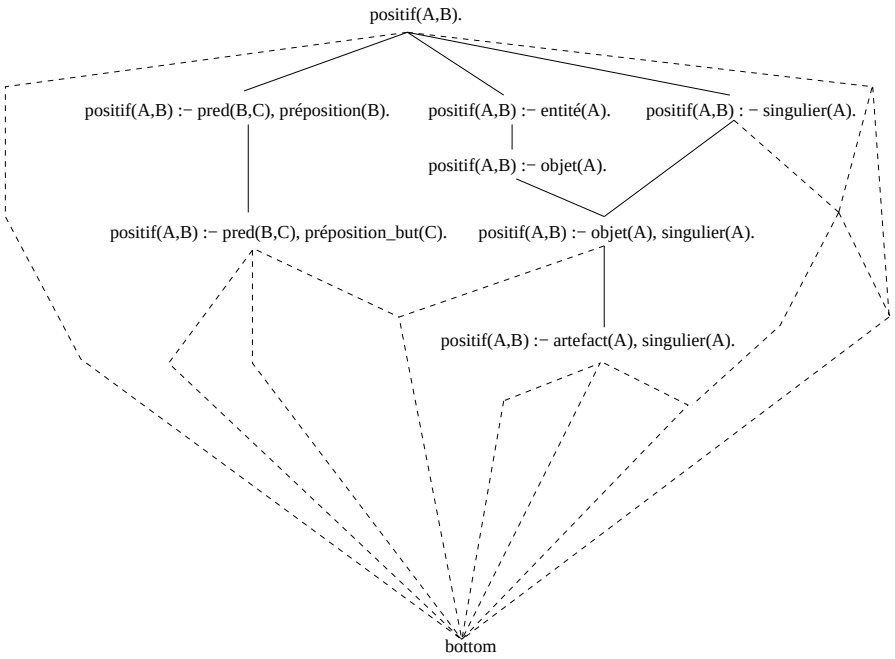

Figure 2. *Exemple de treillis de recherche des hypothèses*

tester exhaustivement toutes les clauses pour s'assurer de cette dernière condition ; il est en effet souvent possible d'élaguer le treillis au cours de la recherche. Par exemple, si une clause ne couvre pas assez d'E^+, il est inutile d'explorer ses raffinements qui, de toutes façons, couvriront encore moins d'exemples. De ce fait, l'élagage permet une amélioration importante de la rapidité de l'algorithme.

Un bon opérateur doit donc permettre autant que faire se peut d'élaguer le treillis de la manière la plus achevée possible. Il doit de plus respecter certaines propriétés [LAA 95], dont la complétude (toutes les hypothèses peuvent être atteintes), la non-redondance (il n'y a qu'une façon d'accéder à une hypothèse), etc. L'écriture d'un opérateur de raffinement est donc un point crucial d'efficacité et d'expressivité pour tout problème de PLI.

Dans notre cas, l'opérateur de raffinement s'appuie sur la hiérarchie des classes sémantiques des mots pour parcourir les clauses du plus général au plus spécifique. Cela est illustré dans la figure 2 où la clause positif(A,B) :- entité(A) est raffinée en positif(A,B) :- objet(A). Le parcours structuré de notre espace de recherche nous a ainsi permis des gains importants d'efficacité, ramenant les temps d'apprentissage à ceux nécessaires dans les deux premières expériences, c'est-à-dire à quelques heures.

4.2. *Troisième expérience d'apprentissage*

L'apprentissage mené à l'aide de cet opérateur de raffinement permet d'obtenir des clauses qui sont donc de la forme de la suivante :
positif(A,B) :- pred(A,C), suc(B,C), preposition_maniere(C), evenement(A), participe(B).
qui signifie que tout couple N-V (i) tel que le mot qui précède le N (pred(A,C)) et qui suit le V (suc(B,C)) est une préposition de manière (preposition_maniere(C)), (ii) tel que le N dénote un événement (evenement(A)) et (iii) tel que le V est au participe passé (participe(B)), peut être considéré comme couple qualia.

L'évaluation théorique de l'apprentissage ainsi effectué, en prenant en compte l'intégralité des informations catégorielles et sémantiques des mots entourant les couples et des constituants de ces couples, est menée d'une manière similaire aux deux expériences précédentes. 63 clauses généralisées sont obtenues. Elles couvrent 78 % des exemples positifs et 3,6 % des exemples négatifs. Le coefficient de Pearson est égal à 0,78 ; l'apprentissage est donc moins bon que pour les expériences précédentes. En effet, le taux de couverture des E^+ assez bas semble indiquer que les clauses produites sont moins générales que celles obtenues lors des expériences précédentes. Nous revenons dans la section suivante sur la relative faiblesse de ce résultat et sur son explication.

Tout comme précédemment, la capacité des règles apprises à construire des lexiques de couples qualia est évaluée de manière empirique en se basant sur les sept noms témoins. Comme nous l'espérions, les résultats sont meilleurs que ceux obtenus jusqu'alors. En effet 44 des 66 couples (66,7 %) sont trouvés et 17 sont incorrectement détectés, et cela, en comptant un couple bon dès qu'il est détecté au moins une fois par l'une des règles apprises. Cela donne un coefficient de Pearson de 0,606. Une fois de plus, le compromis taux de rappel/taux de précision est donc meilleur et le nombre de détections nécessaires inférieur au précédent (1 contre 2). Cela met donc en relief la pertinence des clauses produites par l'apprentissage et justifie pleinement l'intuition que les informations sémantiques sur les constituants du couple jouent un rôle important dans la définition du concept de couple qualia.

L'examen des clauses obtenues montre que, comme attendu, les clauses généralisées précisent les propriétés sémantiques du nom et du verbe et mettent au jour des schémas syntactico-sémantiques, très caractéristiques du corpus, qui favorisent une interprétation télique ou agentive. Par exemple, la clause suivante :
positif(A,B) :- pred(A,C), suc(B,C), preposition_maniere(C), evenement(A), participe(B).
illustrée par les séquences *signalée par l'allumage, alimentée par la génération*, correspond à une structure passive de type télique dans laquelle le type sémantique de l'instrument (événement) est précisé. Une deuxième clause :
positif(A,B) :- pred(A,C), pred(B,D), suc(B,C), preposition(C), preposition_en(D), entite(A).
illustrée dans les séquences *(comprimer) [...] en poussant sur la porte, (valider) [...] en appuyant sur le poussoir*, met en évidence des schémas dénotant une action typique à produire sur un objet exprimés par la suite « *en* + participe présent + prép ».

D'autres clauses caractérisent avec plus de précision par un filtrage sémantique certaines des relations mises au jour aux étapes précédentes, par exemple entre la préposition *sous* et un objet de type *état* (*sous tension*, etc.) ou bien entre un procès et une action physique (*le remplissage s'effectue*). Une autre clause rend aussi compte du fait qu'un verbe et un nom sont liés s'ils apparaissent dans une relation « V det N1 de N2 » où N1 dénote un groupe (par exemple *vidanger le groupe de réservoirs*).

L'apprentissage mené avec toutes les informations sémantiques et catégorielles disponibles sur le contexte environnant du couple et sur le couple lui-même apporte donc une légère amélioration des résultats de détection de couples qualia (on passe d'un coefficient de Pearson de 0,58 à 0,60). La pertinence linguistique des clauses obtenues relève surtout d'une spécification de règles déjà recueillies lors des expériences précédentes. Il est enfin important de noter à ce stade le coût calculatoire relativement élevé dû à l'exploitation de toutes ces informations, notamment les étiquettes sémantiques des noms qui sont très nombreuses. De même, il convient de rappeler, comme dans la deuxième expérience, le coût humain de la mise en place d'un tel étiquetage sémantique des noms qui nécessite l'examen par un expert de tous les mots du corpus[13].

5. Acquisition performante de couples qualia

Dans cette dernière expérience, nous nous intéressons à une approche hybride dictée par un souci de portabilité et d'efficacité. Les exemples sont codés ici de la même manière qu'en section 4 et l'opérateur de raffinement est identique. Nous ôtons simplement du langage d'hypothèses les prédicats correspondant aux étiquettes sémantiques des noms. Ceci signifie que la généralisation des exemples ne peut plus se faire sur les catégories sémantiques des noms apparaissant dans le contexte de couples N-V, ni sur la catégorie sémantique du N. Les informations de nature sémantique sur les verbes, les prépositions et les autres catégories de mots du corpus sont en revanche conservées.

Ce choix s'explique par le fait que contrairement à l'étiquetage catégoriel qui nécessite peu d'intervention humaine et reste relativement portable d'un corpus à l'autre, l'étiquetage sémantique est très coûteux et peu adaptable d'un corpus à un autre, notamment à cause de la construction du lexique des noms qui est la catégorie apportant le plus d'ambiguïtés (voir section 3.1). Nous voulons donc ici confronter notre méthode d'apprentissage à un corpus qui ne comporte que des informations pouvant être ajoutées de manière quasi automatique et peu coûteuse.

L'apprentissage réalisé avec Aleph dans ces conditions donne de bons résultats puisque 93 % des exemples positifs et seulement 3 % des exemples négatifs sont géné-

13. Plusieurs journées de travail (temps de familiarisation avec le corpus et le domaine d'étude inclus) ont été nécessaires à une personne pour réaliser la classification des noms du corpus MATRA CCR.

ralisés, soit un coefficient de Pearson de 0,89. La qualité intrinsèque de l'apprentissage est donc de nouveau comparable aux deux premières expériences.

La validation empirique donne en revanche un résultat qui peut paraître à première vue surprenant. En effet, en utilisant toujours la même méthodologie (application des clauses sur un extrait du corpus MATRA CCR et comparaison des couples obtenus avec les résultats d'un expert pour les sept noms significatifs), nous trouvons, en considérant un couple comme qualia dès sa première détection, 48 des 66 couples qualia et seulement 19 incorrects, soit un coefficient de Pearson de 0,637. Les résultats empiriques sont donc meilleurs que ceux obtenus en prenant en compte les informations sémantiques des noms. Cela s'explique simplement par le fait que l'apprentissage de l'expérience précédente a été de moins bonne qualité (Pearson de 0,78 contre 0,89 ici), ce qui s'est donc traduit par la génération de clauses moins intéressantes qu'elles n'auraient pu l'être. Le problème dans la troisième expérience est vraisemblablement dû au manque d'exemples positifs en regard de la précision de description disponible pour l'algorithme de PLI. En effet, l'utilisation des informations sémantiques sur les noms implique d'ajouter au langage d'hypothèses d'inférence de programmes 33 nouveaux mots correspondant aux 33 étiquettes sémantiques des noms. La granularité trop fine des hypothèses qui en résulte conduit l'algorithme à apprendre « par cœur » les E^+ (on parle alors d'*overfitting*, les règles générées reprenant en partie les exemples sans les généraliser) faute de pouvoir trouver des régularités dans les exemples trop peu nombreux. L'utilité des informations sémantiques n'est donc pas remise en cause par cette dernière expérience, mais soulève un problème important d'alimentation en exemples de notre système d'acquisition de couples qualia.

Les clauses obtenues en n'utilisant donc que les informations catégorielles des mots et les informations sémantiques sur les catégories de mots autres que les noms reprennent pour la plupart des schémas linguistiques donnés en section 2 ou en section 3. Une fois de plus, nous notons l'importance du critère de proximité puisque la clause : positif(A,B) :- contigu(A,B)., qui indique qu'un couple N-V est qualia si N et V sont mitoyens dans une phrase, explique à elle seule près de 2 000 exemples. De même, on retrouve les clauses marquant les énumérations, nombreuses dans notre corpus, à l'aide des ponctuations. Par exemple, la clause :
positif(A,B) :- pred(A,C), pred(B,D), suc(A,D), ponctuation(C), deux_points(D).
généralise les constructions déjà vues page 737 du type « *Poser les raccords : joints alu,* **écrous** *:* **serrer**, … ». Enfin, l'intérêt de l'étiquetage sémantique des prépositions apparaît clairement dans les clauses comme :
positif(A,B) :- pred(A,C), pred(B,D), suc(B,C), preposition_maniere(C), conjonction_coordination(D).
qui détecte grâce à la préposition de manière – marqueur privilégié d'une relation télique – ce type de couple : « *Poser les couvre-joints et* **fixer** *avec les* **vis** *sans serrer.* ».

Cette dernière expérience donne donc, comme les trois précédentes, de très bons résultats, à la fois en termes de production de règles linguistiquement pertinentes et en termes de construction de lexiques de couples qualia. De plus, elle reste relativement

portable d'un corpus à l'autre puisque les informations apportées au corpus peuvent l'être de manière quasi automatique.

6. Conclusions et perspectives

Notre objectif initial était double : d'une part, étudier la faisabilité d'une acquisition automatique de certains éléments du Lexique Génératif sur corpus, à l'aide d'une méthode d'apprentissage qui explique ce qui distingue les couples qualia des autres ; d'autre part, mettre au point une telle méthode qui soit à la fois fiable et peu coûteuse. Le tableau 2 résume les résultats obtenus par nos quatre expériences d'apprentissage. La première colonne indique le coefficient de Pearson de la phase d'apprentissage (c'est-à-dire l'évaluation théorique de la méthode d'apprentissage), la deuxième donne le coefficient de Pearson de la validation sur corpus et la dernière fournit le nombre nécessaire de détections par les clauses pour considérer un couple comme qualia.

	Qualité de l'apprentissage	Validation empirique	Détections nécessaires
Informations catégorielles (voir section 2)	0,84	0,52	6
Informations sémantiques et catégorielles sur le contexte (voir section 3)	0,91	0,58	2
Informations catégorielles et sémantiques (voir section 4)	0,78	0,61	1
Approche hybride (voir section 5)	0,89	0,64	1

Tableau 2. *Résultats des quatre expériences*

Ces quatre expériences, mises en perspective les unes par rapport aux autres, permettent de tirer plusieurs conclusions.

Tout d'abord, l'apprentissage de règles pour la construction de lexiques de couples qualia *est faisable* et donne des résultats déjà corrects en se basant uniquement sur les informations catégorielles des mots. Ensuite, ces résultats sont nettement améliorés par l'apport d'informations sémantiques sur le contexte du couple et aussi sur les constituants même du couple. Enfin, une approche hybride consistant à utiliser les informations sémantiques de toutes les catégories de mots sauf des noms donne elle aussi de très bons résultats.

Les règles obtenues sont, nous l'avons vu, très dépendantes du corpus et plus encore de la langue (la souplesse de notre méthode nous permettant d'envisager facilement le traitement de corpus d'autres langues que le français). L'apprentissage de ces règles nécessite donc d'être reconduit sur tout ensemble de textes d'un nouveau

domaine. Dès lors, il est important de considérer le *coût* (c'est-à-dire en particulier le temps) nécessaire au portage de la procédure dans son ensemble d'un corpus à un autre. La première méthode est peu coûteuse, mais elle gagne à être étoffée à l'aide de connaissances sémantiques. L'utilisation des informations sémantiques, et en particulier celles des noms, engendre cependant des coûts supplémentaires importants qui sont de deux ordres, et qui soulèvent donc la question de la portabilité de notre méthode d'apprentissage. Le premier est dû à la mise au point de la classification des noms du corpus d'étude qui sert à la construction du lexique des noms utilisé lors de l'étiquetage sémantique. Ce lexique, évidemment propre à chaque corpus, est le plus long à construire et aussi celui qui porte le plus d'ambiguïtés par rapport aux lexiques des autres catégories de mots. Le second type de coût est lié au fait que l'utilisation des informations sémantiques implique, à cause de la richesse du langage d'hypothèses engendré, de construire un nombre considérable d'exemples. Or cette tâche est l'une des plus coûteuses du processus puisqu'il s'agit d'annoter à la main les occurrences de couples au sein des phrases du corpus. Toutefois, en considérant les excellents résultats et la bonne portabilité de l'approche décrite en section 5, et en comparant ses performances avec celle présentée en section 4, il paraît parfaitement raisonnable d'utiliser la technique hybride pour une simple tâche de construction de lexiques de couple qualia. Nos travaux étant guidés par l'utilisation en recherche d'information de ces couples de type qualia[14] – et donc par la construction de manière quasi automatique de ces lexiques – nous privilégions évidemment cette dernière approche. Nous envisageons d'ailleurs d'étendre cette méthode à l'acquisition de couples nom-nom de type qualia dans de futurs travaux (par exemple *livre lecture*, etc.).

Une autre perspective consiste à exploiter le caractère explicatif de notre méthode d'apprentissage pour découvrir les fondements linguistiques de la notion de rôle qualia. Nous avons déjà signalé au cours des diverses expériences la réelle signification linguistique des règles apprises. Nous avons également comparé les clauses que nous avons acquises avec des explorations manuelles du corpus ayant pour but de découvrir les structures exprimant les relations qualia [GAL 00]. Cette comparaison montre que les principales structures mises en évidence sont correctement apprises par notre méthode. Certaines structures valides ne sont cependant pas retrouvées, en particulier quand les marqueurs sont des expressions polylexicales telles que « avoir pour but de », « être utilisé pour », « avoir pour fonction de », etc. En revanche, notre technique d'apprentissage suggère l'importance d'un critère de position (tel « V doit se trouver au début de la phrase ») qui était resté inaperçu dans les observations faites dans [GAL 00]. Il est possible dans ce but de poursuivre les expérimentations menées en s'inspirant de ce qui a été fait en section 4, puisque c'est cette expérience qui fournit les clauses explicatives les plus spécialisées. Il conviendra alors de fournir plus d'exemples et d'améliorer encore l'efficacité de la phase d'apprentissage à travers l'écriture d'un opérateur de raffinement performant. Une étape supplémentaire de post-traitement des clauses obtenues sera alors sans doute nécessaire. Il est en effet

14. Cet objectif d'utilisation des couples N-V en recherche d'information explique d'ailleurs le fait que nous n'ayons pas cherché jusqu'ici à différencier les divers rôles qualia parmi les paires N-V apprises. La méthode que nous avons exposée ici permet cependant d'envisager cette tâche.

probable que l'accroissement du nombre d'exemples et l'extension du langage d'hypothèses conduisent le processus d'apprentissage à générer un trop grand nombre de règles pour verbaliser de manière compacte la théorie du LG.

7. Bibliographie

[AGA 95] AGARWAL R., Semantic Feature Extraction from Technical Texts with Limited Human Intervention, PhD thesis, Mississippi State University, États-Unis, 1995.

[ARM 95] ARMSTRONG S., BOUILLON P., ROBERT G., Tagger Overview, Rapport, 1995, ISSCO, (http://www.issco.unige.ch/staff/robert/tatoo/ tagger.html).

[ARM 96] ARMSTRONG S., « MULTEXT: Multilingual Text Tools and Corpora », FELDWEG H., HINRICHS W., Eds., *Lexikon und Text*, p. 107-119, Tübingen: Niemeyer, 1996.

[BOU 97] BOUAUD J., HABERT B., NAZARENKO A., ZWEIGENBAUM P., « Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation avec deux modélisations conceptuelles », *Ingénierie de la Connaissance*, Roscoff, France, 1997.

[BOU 00] BOUILLON P., FABRE C., SÉBILLOT P., JACQMIN L., « Apprentissage de ressources lexicales pour l'extension de requêtes », *TAL (traitement automatique des langues), numéro spécial Traitement automatique des langues pour la recherche d'information*, vol. 41, n° 2, 2000, p. 367-393.

[BOU 01a] BOUILLON P., BUSA F., *Generativity in the Lexicon*, CUP: Cambridge, 2001.

[BOU 01b] BOUILLON P., CLAVEAU V., FABRE C., SÉBILLOT P., « Using Part-of-Speech and Semantic Tagging for the Corpus-Based Learning of Qualia Structure Elements », *First International Workshop on Generative Approaches to the Lexicon, GL'2001*, Genève, Suisse, 2001.

[BRI 97] BRISCOE T., CARROLL J., « Automatic Extraction of Subcategorisation from Corpora », *5th ACL Conference on Applied Natural Language Processing*, Washington, États-Unis, 1997.

[BUI 97] BUITELAAR P., « A Lexicon for Underspecified Semantic Tagging », *ANLP'97 Workshop on Tagging text with Lexical Semantics*, Washington, États-Unis, 1997.

[CEU 96] CEUSTERS W., SPYNS P., DEMOOR G., MARTIN W., *Tagging of Medical Texts: The Multi-TALE Project*, Amsterdam: IOS Press, 1996.

[FAB 96] FABRE C., Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations, Thèse, Université de Rennes 1, 1996.

[FAB 99] FABRE C., SÉBILLOT P., « Semantic Interpretation of Binominal Sequences and Information Retrieval », *International ICSC Congress on Computational Intelligence: Methods and Applications, CIMA'99, Symposium on Advances in Intelligent Data Analysis, AIDA'99*, Rochester, N.Y., États-Unis, 1999.

[FAU 99] FAURE D., NÉDELLEC C., « Knowledge Acquisition of Predicate Argument Structures from Technical Texts using Machine Learning: the System ASIUM », FENSEL D., STUDER R., Eds., *11th European Workshop EKAW'99*, Dagstuhl, Allemagne, 1999, Springer-Verlag.

[FEL 98] FELLBAUM C., Ed., *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998.

[GAL 00] ÉDITH GALY, Repérer en corpus les associations sémantiques privilégiées entre le nom et le verbe : le cas de la fonction dénotée par le nom, Mémoire de Maîtrise, Université de Toulouse-Le Mirail, France, 2000.

[GRE 94a] GREFENSTETTE G., « Corpus-Derived First, Second and Third-Order Word Affinities », *EURALEX'94*, Amsterdam, Pays-Bas, 1994.

[GRE 94b] GREFENSTETTE G., *Explorations in Automatic Thesaurus Discovery*, Dordrecht, Kluwer Academic Publishers, 1994.

[GRE 97] GREFENSTETTE G., « SQLET: Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text », *Recherche d'Informations Assistée par Ordinateur, RIAO'97*, Montréal, Québec, Canada, 1997, McGill-University.

[HAB 97] HABERT B., NAZARENKO A., SALEM A., *Les linguistiques de corpus*, Armand Collin/Masson, Paris, 1997.

[HAR 89] HARRIS Z., GOTTFRIED M., RYCKMAN T., MATTICK(JR) P., DALADIER A., HARRIS T. N., HARRIS S., *The Form of Information in Science, Analysis of Immunology Sublanguage*, Kluwer Academic Publisher, Dordrecht, 1989.

[HEA 92] HEARST M. A., « Automatic Acquisition of Hyponyms from Large Text Corpora », *15th International Conference on Computational Linguistics, COLING'92*, Nantes, France, 1992.

[LAA 95] VAN DER LAAG P., An analysis of refinement operators in inductive logic programming, PhD thesis, Erasmus Universiteit, Rotterdam, Pays-Bas, 1995.

[MIT 97] MITCHELL T. M., *Machine Learning*, McGraw-Hill, 1997.

[MOR 97] MORIN E., « Extraction de liens sémantiques entre termes dans des corpus de textes techniques : application à l'hyponymie », *Traitement Automatique des Langues Naturelles, TALN'97*, Grenoble, France, 1997.

[MUG 94] MUGGLETON S., DE-RAEDT L., « Inductive Logic Programming: Theory and Methods », *Journal of Logic Programming*, vol. 19-20, 1994, p. 629-679.

[MUG 95] MUGGLETON S., « Inverse Entailment and Progol », *New Generation Computing*, vol. 13, n° 3-4, 1995, p. 245-286.

[PIC 97] PICHON R., SÉBILLOT P., Acquisition automatique d'informations lexicales à partir de corpus : un bilan, Rapport de recherche n° 3321, 1997, INRIA, Rennes.

[PIC 00] PICHON R., SÉBILLOT P., « From Corpus to Lexicon: from Contexts to Semantic Features », LEWANDOWSKA-TOMASZCZYK B., MELIA P. J., Eds., *PALC'99: Practical Applications in Language Corpora*, Łódź studies in Language, Peter Lang, 2000.

[PUS 93] PUSTEJOVSKY J., BERGLER S., ANICK P., « Lexical Semantic Techniques for Corpus Analysis », *Computational Linguistics*, vol. 19, n° 2, 1993.

[PUS 95] PUSTEJOVSKY J., *The Generative Lexicon*, Cambridge: MIT Press, 1995.

[ROB 98] ROBERTS S., VAN-LAER W., JACOBS N., MUGGLETON S., BROUGHTON J., « A Comparison of ILP and Propositional Systems on Propositional Data », *8th International Workshop on Inductive Logic Programming, ILP-98*, Berlin, Allemagne, 1998, Springer-Verlag, LNAI 1446.

[SÉB 00] SÉBILLOT P., BOUILLON P., FABRE C., « Inductive Logic Programming for Corpus-Based Acquisition of Semantic Lexicons », *Learning Language in Logic, LLL-2000*, Lisbonne, Portugal, 2000.

[SHA 81] SHAPIRO E. Y., Inductive Inference of Theories from Facts, Research report n° 624, 1981, Department of Computer Science, Yale University, New Haven.

[SME 99] SMEATON A. F., « Using NLP or NLP resources for Information Retrieval Tasks », STRZALKOWSKI T., Ed., *Natural Language Information Retrieval*, p. 99-111, Kluwer Academic Publishers, 1999.

[VOO 94] VOORHEES E. M., « Query Expansion using Lexical-Semantic Relations », *ACM SIGIR'94*, Dublin, Irlande, 1994.

[WER 96] WERMTER S., RILOFF E., SCHELER G., Eds., *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Lecture Notes in Computer Science, vol. 1040, Springer-Verlag, 1996.

[WIL 96] WILKS Y., STEVENSON M., The Grammar of Sense: is Word-Sense Tagging much more than Part-of-Speech Tagging?, Rapport, 1996, University of Sheffield, Grande-Bretagne.

[YAR 92] YAROWSKY D., « Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora », *15th International Conference on Computational Linguistics, COLING'92*, Nantes, France, 1992.