Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbe

Vincent Claveau* — **Pascale Sébillot****

* *OLST - Université de Montréal*
C.P. 6128, succ. Centre-Ville
Montréal, QC, H2J 2Y2, Canada
Vincent.Claveau@UMontreal.CA

** *IRISA - Université de Rennes 1*
Campus de Beaulieu
F-35042 Rennes cedex
Pascale.Sebillot@irisa.fr

RÉSUMÉ. *Dans le modèle du Lexique génératif (LG) [PUS 95], certaines propriétés sémantiques des noms sont exprimées à l'aide de verbes. De tels couples nom-verbe présentent un intérêt applicatif notamment en recherche d'information. Leur acquisition sur corpus constitue donc un enjeu, mais la découverte des patrons qui les définissent en contexte est également importante pour la compréhension même du modèle du LG. Cet article présente deux techniques entièrement automatiques répondant à ce double besoin d'extraction de couples et de patrons contextuels. Elles combinent pour ce faire deux approches d'acquisition — les approches statistique et symbolique — en conservant les avantages propres à chacune d'entre elles : robustesse et automaticité des premières, qualité et expressivité des résultats des secondes.*

ABSTRACT. *In the Generative Lexicon (GL) framework [PUS 95], some semantic properties of nouns are expressed with the help of verbs. These noun-verb pairs are relevant in various domains, such as Information Retrieval. Their corpus-based acquisition is thus an interesting issue; moreover, discovering the contextual patterns in which these pairs occur can help to understand the GL model. This paper presents two related and fully automated techniques that allow us to acquire from a corpus both noun-verb pairs, and semantic and morpho-syntactic patterns. These techniques combine two different acquisition approaches—the statistical one and the symbolic one—and keep advantages of each approach: robustness and automation of statistical methods, quality of the results and expressiveness of symbolic ones.*

MOTS-CLÉS : *Sémantique lexicale, extraction de patrons contextuels, Lexique génératif, programmation logique inductive, bootstrapping, apprentissage artificiel semi-supervisé.*

KEYWORDS: *Lexical semantics, contextual pattern extraction, Generative Lexicon, inductive logic programming, bootstrapping, semi-supervised machine learning.*

1. Introduction

Le Lexique génératif est un modèle de lexique développé par J. Pustejovsky [PUS 95, BOU 01a] dans lequel les entrées lexicales sont composées de quatre structures. L'une d'entre elles, appelée *structure des qualia*, donne accès à différentes propriétés sémantiques du mot à travers quatre formules prédicatives typées, essentiellement verbales : les rôles qualia (télique, agentif, constitutif et formel). La structure des qualia du nom *livre* contient par exemple les prédicats *lire* (le rôle télique, indiquant le but ou la fonction), *écrire* (le rôle agentif, indiquant le mode de création) et *contenir (de l'information)* (le rôle formel, indiquant la relation que le mot entretient avec les classes sémantiques dont il hérite). Chaque entrée lexicale, et plus spécialement celles des noms (N), est ainsi liée à des verbes (V) indispensables à leur interprétation syntaxique et sémantique en contexte ; ce sont ces paires N-V, dans lesquelles le V appartient à la structure des qualia du N et que nous nommons *couples qualia* par la suite, que nous nous proposons d'acquérir automatiquement sur corpus.

C. Fabre *et al.* [FAB 99, FAB 96] ont montré l'intérêt des relations N-V qualia pour permettre l'interprétation des structures complexes binominales, dans lesquelles une part essentielle du contenu renvoie à des informations de nature prédicative exprimables à l'aide d'un verbe (*panier à pain - stocker, entreposer*). L'exploitation du lien nom-verbe donne également accès à certaines variantes de termes très utiles quoique peu usitées jusqu'à présent dans le domaine de la recherche d'information [GRE 97, CLA 04b]. Elle permet par exemple des reformulations de requêtes du type *magasin de disque* $\Rightarrow$ *vendre des disques* grâce au couple qualia *magasin-vendre*. Bien que leur intérêt ait été montré pour ce type de tâche [BOU 00b, PUS 97], l'absence de telles ressources lexicales en empêche toutefois l'utilisation à grande échelle.

Le formalisme proposé par Pustejovsky, développé uniquement dans un cadre théorique, ne suggère aucune méthode de construction des entrées lexicales. L'acquisition sur corpus d'éléments du Lexique génératif tels que les couples qualia est donc en soi un enjeu intéressant, mais, parallèlement à cette tâche d'extraction, notre objectif est aussi de fournir une aide à la verbalisation du concept de rôle qualia par l'obtention de règles d'extraction linguistiquement pertinentes. Par ailleurs, les prédicats jouant différents rôles qualia pour un nom donné sont propres à chaque domaine ; il paraît donc essentiel de développer une méthode d'acquisition sur corpus aisément portable d'un texte à un autre, pour laquelle la capacité d'adaptation et le degré d'automatisation seront des critères d'évaluation au même titre que la qualité des résultats qu'elle produit.

Parmi les rares travaux entrepris jusqu'ici pour la construction de structures des qualia, [PUS 93] propose d'acquérir les éléments de ces structures à partir d'un texte étiqueté syntaxiquement en utilisant une extraction statistique de cooccurrences couplée à un jeu d'heuristiques sous forme de patrons syntaxiques. Malheureusement, cette étude n'est pas clairement évaluée, et la question de la portabilité de la méthode d'un texte à un autre n'est pas abordée, notamment en ce qui concerne l'étiquetage syntaxique et les heuristiques employées.

Pour notre part, des travaux antérieurs [CLA 03b] nous ont permis de mettre au point et d'évaluer une technique d'extraction de couples N-V qualia basée sur une méthode d'apprentissage symbolique supervisé : la programmation logique inductive (PLI) [MUG 94b]. La PLI sert à générer un classifieur (un ensemble de règles Prolog) de manière supervisée, c'est-à-dire à l'aide d'exemples de couples qualia en contexte et de contre-exemples (couples N-V non-qualia au sein d'une phrase). Les règles obtenues sont ensuite utilisées comme patrons d'extraction. Les résultats de cette technique sont de bonne qualité, aussi bien pour la tâche d'extraction des couples que pour la pertinence linguistique des règles générées. En effet, les patrons morphosyntaxiques et sémantiques obtenus fournissent un support interprétable permettant la définition du concept même de rôle qualia. Cependant, ces patrons étant *a priori* propres à chaque corpus, il est nécessaire de reconduire la phase d'apprentissage par PLI pour tout nouveau texte. Cette approche supervisée présente donc l'inconvénient intrinsèque de nécessiter la construction d'un jeu d'exemples et de contre-exemples de couples qualia propres au corpus à traiter. Cette phase de supervision, essentiellement manuelle et exigeant l'aide d'un expert, est très coûteuse, et conduit notre technique d'acquisition actuelle à ne pas répondre entièrement à notre souci de portabilité et d'automaticité.

Dans cet article, nous proposons deux variantes de notre méthode d'apprentissage de couples N-V qualia qui remédient à ce problème et remplissent ainsi les différentes exigences citées précédemment : bonne qualité des résultats, interprétabilité linguistique et automaticité du processus. Nous rejoignons en cela certains travaux effectués en extraction d'informations [RIL 99] réduisant l'intervention humaine à l'apport de quelques données (*seed-words*) en début de processus ; notre système d'acquisition se veut quant à lui entièrement automatique et sans aucun appui humain. Pour ce faire, en conjonction de notre approche symbolique supervisée, nous utilisons de deux façons différentes une technique reposant sur une approche différente de l'extraction : l'approche statistique. Les systèmes d'extraction mixtes résultants sont entièrement automatiques et ne nécessitent plus de fournir manuellement des exemples de couples qualia à l'algorithme de PLI. Ces deux variantes non supervisées obtiennent des résultats similaires à la technique originale.

Après un positionnement de nos travaux d'acquisition de couples N-V qualia par rapport à d'autres études menées dans le cadre de l'extraction de relations lexicales à partir de corpus (section 2), nous décrivons en section 3 notre première approche entièrement symbolique d'apprentissage de paires qualia, et présentons une évaluation de ses résultats. La section 4 propose quant à elle les deux variantes de cette technique combinant extractions symbolique et statistique ; les performances de ces deux nouveaux systèmes mixtes d'acquisition y sont examinées et comparées à celles de la version initiale. Enfin nous concluons en revenant sur les qualités et les inconvénients des systèmes proposés et présentons quelques perspectives à ces travaux en dernière partie.

2. Positionnement par rapport aux travaux d'acquisition de relations lexicales

Le domaine de l'acquisition de relations lexicales à partir de corpus a donné lieu à un nombre conséquent de travaux au cours de ces dernières années (*cf.* par exemple [PIC 97, CLA 03a] pour une vue du domaine). Les différentes techniques d'acquisition mises au point dans ce cadre peuvent être vues comme fondées sur la notion de *classifieur*, en ce sens qu'elles produisent des méthodes permettant de classer les divers couples (ou n-uplets) d'unités lexicales observés comme respectant ou non la relation cible.

Ces classifieurs peuvent être regroupés en différentes familles selon les attributs qu'ils utilisent pour repérer les éléments suivant la relation recherchée. On distingue usuellement les travaux qui se basent sur l'aspect fréquentiel du corpus duquel ils extraient les n-uplets en relation, et ceux qui exploitent des indices structurels pour détecter les éléments liés, c'est-à-dire qui suivent une approche symbolique[1]. Nous présentons successivement ces deux familles, s'appuyant respectivement sur des indices numériques et structurels, avant de conclure en explicitant notre positionnement.

2.1. *Acquisition à partir d'indices numériques*

Les études du premier type sont de loin les plus nombreuses (*cf.* [HAB 97, GRE 94b]). Elles visent à extraire des informations syntagmatiques et paradigmatiques sur des unités lexicales en étudiant respectivement les mots qui apparaissent dans la même fenêtre ou le même contexte syntaxique 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). Reposant sur des informations fréquentielles, l'extraction d'éléments respectant la relation recherchée est réalisée au niveau du corpus pris dans son ensemble. Des indices statistiques d'association [MAN 99] sont un outil couramment utilisé pour mettre au jour les relations syntagmatiques en pointant les mots qui cooccurrent dans une zone de texte de manière statistiquement significative. Les travaux menés sur les relations paradigmatiques se placent quant à eux volontiers dans le cadre de la linguistique harrissienne [HAR 89] qui pose l'hypothèse qu'il est possible de mettre en évidence, à partir d'une analyse distributionnelle de contextes rendus élémentaires, les classes de concepts et les relations d'un sous-langage lié à un domaine d'activité. Parmi ces nombreuses recherches, [BRI 97] tente par exemple d'apprendre automatiquement des structures argumentales et des restrictions sélectionnelles ; [GRE 95] acquiert des verbes supports de nominalisations ; [AGA 95, BOU 97, BOU 02b] construisent des

1. En pratique, plusieurs méthodes d'acquisition existantes se fondent à la fois sur les deux aspects, numérique et symbolique, pour acquérir des éléments en relation lexicale. Il est également important de noter que l'aspect des données exploité par le classifieur ne préjuge en rien de la technique utilisée pour produire le classifieur ; des techniques d'apprentissage numérique peuvent par exemple produire des classifieurs symboliques.

classes sémantiques ; [GRE 94b] vise de son côté l'obtention de représentations lexicales sémantiques plus complètes.

2.2. *Acquisition à partir d'indices structurels*

Les méthodes symboliques d'extraction de relations lexicales s'appuient quant à elles sur des indices collectés sur le contexte d'une occurrence de mots en relation pour décider de son acquisition ou non ; le classifieur symbolique est donc souvent un ensemble de règles s'appuyant sur des indices lexicaux, morphologiques, catégoriels, syntaxiques ou autres. Ces techniques peuvent elles-mêmes se classer en deux grandes familles : les approches linguistiques dans lesquelles des indices structurels donnés *a priori* (par une analyse linguistique par exemple) sont exploités, et les approches basées sur une notion d'apprentissage (artificiel ou non).

2.2.1. *Approche linguistique*

Parmi les premières, on peut citer le système SEEK [JOU 95] qui utilise des règles d'exploration contextuelles, reposant sur l'identification de marqueurs linguistiques et construites manuellement, pour détecter des relations binaires variées (inclusion, identification, tout à partie...), ou le système COATIS [GAR 00] qui se focalise sur la relation de causalité, voire FASTR de C. Jacquemin [JAC 01], en percevant les variantes morphosyntaxiques de termes comme des relations d'équivalence.

2.2.2. *Approche par apprentissage*

L'acquisition de relations sémantiques par apprentissage de patrons d'extraction lexico-syntaxiques est le point commun des recherches du second type. La plupart d'entre elles consistent à identifier dans un corpus des marqueurs ou indices d'une relation sémantique à partir d'un petit ensemble d'exemples, et à les réutiliser ensuite pour extraire de nouvelles unités en relation. C'est l'approche initiée par M. Hearst [HEA 92, HEA 98] pour acquérir des liens d'hyperonymie, qui se formalise en cinq étapes :

1) choisir une relation cible $\mathcal{R}$;

2) réunir une liste de paires en relation $\mathcal{R}$ (par exemple les extraire d'un thésaurus, d'une base de connaissances) ;

3) retrouver les phrases du corpus contenant ces paires et enregistrer leurs contextes lexical et syntaxique ;

4) trouver les points communs entre ces contextes et supposer que cela forme un schéma lexico-syntaxique de $\mathcal{R}$;

5) appliquer les schémas pour obtenir de nouvelles paires et retourner en 3.

Contrairement aux recherches citées au paragraphe précédent, les marqueurs d'une relation sont donc ici issus d'une analyse d'exemples et non d'une connaissance linguistique *a priori*. Dans les travaux de M. Hearst, la phase 4 est entièrement manuelle.

Dans son système PROMÉTHÉE, E. Morin [MOR 99] en propose une automatisation basée sur un calcul de similarité [MOR 98] entre les contextes lexico-syntaxiques d'occurrences de paires. Des classes de contextes lexico-syntaxiques sont ainsi constituées, d'où émergent des schémas représentatifs obtenus par généralisation d'un des contextes de chaque classe en supprimant tous les attributs non communs aux autres contextes de la même classe.

2.3. *Positionnement*

Comme nous l'avons indiqué en introduction, nos objectifs particuliers, à savoir développer un système d'acquisition de paires N-V qualia qui soit portable, produise de bons résultats, linguistiquement interprétables afin d'expliciter les notions de rôles qualia, nous ont naturellement conduits à nous positionner dans le cadre des approches structurelles, et plus précisément celui de l'apprentissage artificiel symbolique de patrons.

Les méthodes d'acquisition à partir d'indices numériques, si elles sont portables et automatiques, souffrent en effet d'un manque d'interprétabilité. Il est souvent difficile de comprendre pourquoi un couple d'éléments cooccurrants a été retenu et pas un autre, le seul indice fourni à ce sujet étant généralement un score statistique. Ces méthodes n'offrent donc aucun retour sur la définition de l'information recherchée. Les méthodes symboliques suivant une approche linguistique font quant à elles l'hypothèse que les relations qu'elles décrivent par des indices prédéfinis sont suffisamment génériques pour ne pas dépendre d'un domaine particulier. Or certaines expériences [JOU 97] montrent que cette affirmation est difficilement tenable. De plus, dans notre cas particulier, outre le fait que les liens qualia soient portés par des schémas variables d'un corpus à l'autre (ce que nos expériences confirment), nous n'avons aucun indice *a priori* sur d'éventuels marqueurs de la présence de liens qualia.

Notre positionnement est réalisé en pratique par l'utilisation d'une technique éprouvée d'apprentissage supervisé : la programmation logique inductive qui permet la production (on parle alors d'inférence) de patrons décrivant et définissant les relations qualia à partir d'exemples. Par rapport aux travaux d'E. Morin, ce placement dans un cadre d'apprentissage artificiel formel nous permet de donner une assise théorique à la notion de généralisation à partir d'exemples. Nous présentons dans la section suivante ASARES[2], notre système d'acquisition de couples qualia, qui répond à deux des trois critères (qualité et interprétabilité des résultats) que nous nous sommes fixés, avant d'expliciter les moyens que nous avons mis en œuvre pour atteindre le dernier.

2. Acronyme pour Acquisition Symbolique Automatique de REssources Sémantiques.

3. Acquisition symbolique de couples qualia : le système ASARES

Cette section décrit dans un premier temps le corpus utilisé et les différents étiquetages qu'il a subis avant son utilisation par notre système d'acquisition symbolique ASARES. Nous présentons ensuite la technique d'apprentissage employée : la programmation logique inductive. Son utilisation dans notre cadre d'extraction de couples qualia est détaillée en troisième partie. Nous terminons par un examen des résultats obtenus en utilisant ASARES sur le corpus.

3.1. *Corpus et étiquetages*

Le corpus utilisé lors de nos expérimentations est un manuel de maintenance d'hélicoptères, en français, qui nous a été fourni par MATRA-CCR Aérospatiale. Il contient environ 104 000 mots, soit une taille de près de 700 Koctets. Ce corpus technique présente des caractéristiques se prêtant bien à notre tâche d'acquisition : il est très homogène (aussi bien pour ses structures syntaxiques que pour son vocabulaire) ; il contient également de nombreux termes concrets apparaissant fréquemment au sein des phrases avec des verbes indiquant leur rôle télique ou agentif.

Ce corpus a tout d'abord subi un étiquetage catégoriel automatique. À l'aide des outils développés dans le projet MULTEXT [ARM 96], il a donc été segmenté en phrases et en mots, puis lemmatisé et analysé, et enfin désambiguïsé avec TATOO[3], un outil basé sur des chaînes de Markov cachées. Chaque mot a ainsi reçu une étiquette indiquant sa catégorie morphosyntaxique, son genre, son nombre. La précision de cet étiquetage, évaluée à l'aide d'un extrait de 4 000 mots étiquetés à la main est très bonne : moins de 2 % d'erreurs ont été détectées.

Suivant la méthode exposée dans [BOU 00a], un étiquetage sémantique du corpus a également été réalisé. Il est effectué sur le corpus étiqueté catégoriellement, et bénéficie ainsi de la désambiguïsation des mots polyfonctionnels tels que *règle* qui peut être à la fois un verbe à l'indicatif et un nom [WIL 96]. Il consiste en trois étapes. La première, la plus coûteuse du processus parce que principalement manuelle, est de constituer un lexique associant à chaque mot du corpus sa ou ses étiquettes sémantiques. La deuxième étape concerne la projection de ces étiquettes sur le corpus. Enfin, la dernière consiste à désambiguïser les mots ayant reçu plusieurs étiquettes. L'hypothèse originale de cette approche est de supposer que ces ambiguïtés peuvent être résolues, comme dans le cas des étiquettes morpho-syntaxiques, par des techniques de chaînes de Markov cachées. Pour composer le jeu d'étiquettes sémantiques employé, les classes les plus génériques de WORDNET [FEL 98] ont été utilisées et adaptées à notre corpus : les classes non pertinentes (pour notre corpus) ont été supprimées, et pour les classes trop larges, une granularité plus fine a été choisie. Pour les noms, nous obtenons par exemple 33 classes, organisées hiérarchiquement comme indiqué en figure 1 (les classes non utilisées pour l'étiquetage sont en italiques, les

3. Disponible à l'URL http ://www.issco.unige.ch/staff/robert/tatoo/tatoo.html.

étiquettes effectivement employées sont entre parenthèses). Une description plus détaillée du processus d'étiquetage sémantique et du jeu d'étiquettes est donnée dans [BOU 01b, CLA 01]. Le taux d'erreurs détectées, mesuré à l'aide d'un extrait du corpus de 6 000 mots étiquetés à la main, est là encore très faible : 85 % des ambiguïtés sont correctement résolues, soit une précision totale de 98.82 %.

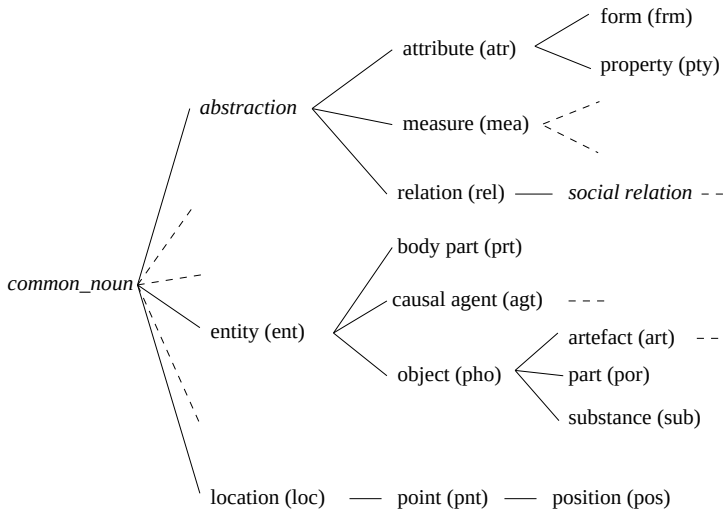

Figure 1. *Extrait de la hiérarchie des classes sémantiques des noms*

3.2. *La programmation logique inductive*

Depuis quelques années, l'utilisation de méthodes d'apprentissage artificiel symbolique pour des tâches relevant du traitement automatique des langues s'est développée. Parmi ces méthodes, la programmation logique inductive (PLI) [MUG 94b], grâce à son expressivité et sa souplesse d'utilisation, a été appliquée à des problèmes aussi divers que l'étiquetage morphosyntaxique, la construction d'analyseurs syntaxiques ou encore l'interrogation de bases de données en langage naturel (voir par exemple [CUS 00] pour un panorama de travaux dans ce domaine). Nous présentons ici succinctement certains principes très généraux de la PLI (se reporter à [CLA 03b] pour une description plus détaillée de cette technique), ainsi que son intérêt particulier pour notre problématique d'extraction de couples qualia.

3.2.1. *Généralités et notations*

La PLI se situe à la croisée de la programmation logique et de l'apprentissage artificiel. Elle est utilisée pour produire (inférer) des règles générales ou hypothèses (sous forme de clauses de Horn) expliquant un concept à partir d'exemples et de

contre-exemples de ce concept et d'un ensemble d'informations préexistantes appelé *background knowledge*. Les règles sont obtenues par généralisation des exemples ; les contre-exemples servent à empêcher une généralisation excessive en vérifiant que les règles produites n'en recouvrent aucun (ou très peu, un peu de *bruit* pouvant être autorisé). L'ensemble des exemples positifs est traditionnellement noté E^+, l'ensemble des contre-exemples ou exemples négatifs, E^-, le *background knowledge*, $\mathcal{B}$, et la PLI cherche à induire un classifieur sous la forme d'un ensemble d'hypothèses H. Un certain nombre de conditions sont imposées sur les données d'apprentissage (où le symbole $\models$ représente l'implication) :

– la consistance (ou satisfiabilité) *a priori*, qui est assurée si les exemples négatifs ne sont pas en contradiction avec les connaissances données dans le *background knowledge*, ce qui se note : $\mathcal{B} \wedge E^- \not\models \square$ ($\square$ signifie faux) ;

– la nécessité *a priori*, qui est la traduction du besoin d'une connaissance autre que $\mathcal{B}$, d'un ensemble d'hypothèses, pour expliquer les exemples positifs, ce que l'on traduit par $\mathcal{B} \not\models E^+$.

Sous ces conditions, l'algorithme de PLI cherche l'ensemble H tel que les conditions suivantes soient satisfaites :

– la consistance (ou satisfiabilité) *a posteriori*, qui impose qu'aucune contradiction ne soit trouvée entre $\mathcal{B}$, H et E^-, ce que l'on note : $\mathcal{B} \wedge H \wedge E^- \not\models \square$;

– la complétude, qui consiste à s'assurer que l'hypothèse, combinée au *background knowledge*, permet bien d'expliquer tous les exemples positifs, soit $\mathcal{B} \wedge H \models E^+$.

L'avantage majeur de la PLI est de permettre l'apprentissage à partir d'exemples relationnels (c'est-à-dire qu'on ne peut pas décrire par un ensemble de couples attribut-valeur) ainsi que l'apprentissage de concepts relationnels, usuellement exprimés en Prolog. C'est cette expressivité à la fois en entrée et en sortie du processus d'apprentissage qui rend cette technique très adaptée pour traiter certains problèmes difficilement exprimables hors de ce cadre relationnel.

Un *langage d'hypothèses* $\mathscr{L}_H$ est également donné à l'algorithme de PLI ; il est utilisé pour définir précisément la forme attendue des règles générées. Ce langage assure ainsi de n'obtenir que des règles bien formées et pertinentes au regard de la tâche d'apprentissage traitée. En fonction de ce langage, l'objectif de la PLI est donc d'inférer des règles qui couvrent (c'est-à-dire expliquent) un maximum d'exemples et aucun contre-exemple (ou très peu selon le bruit autorisé). Bien que restreint par les contraintes exprimées à travers $\mathscr{L}_H$, l'espace de recherche est en général très vaste, voire infini. Toutefois les hypothèses qui le composent peuvent être organisées par une relation de généralité qui permet de le parcourir intelligemment à l'aide d'un opérateur appelé opérateur de raffinement. Lors de ce parcours, le choix d'une hypothèse h est fait selon une fonction de score Sc, qui dépend le plus souvent du nombre d'exemples positifs et négatifs couverts par l'hypothèse. Ces deux ensembles sont respectivement notés E_h^+ et E_h^- et leurs cardinaux $|E_h^+|$ et $|E_h^-|$ dans la suite de cet article.

3.2.2. *Choix de la PLI pour notre application*

Dans notre cas, le concept que nous cherchons à apprendre est la nature qualia d'une paire N-V apparaissant au sein d'une phrase. Nous souhaitons donc obtenir un classifieur permettant, d'une part, de distinguer en contexte une paire N-V qualia d'une non-qualia, mais aussi de refléter des éléments linguistiquement interprétables définissant le concept même de rôle qualia. La PLI, grâce à son aspect explicatif, se révèle un choix bien adapté pour cette tâche puisque les règles générées peuvent ensuite être directement interprétées comme des patrons linguistiques d'extraction et offrent une expressivité propre à la logique des prédicats. Cette expressivité de la PLI est également exploitée pour l'encodage des exemples, c'est-à-dire la description des phrases contenant des couples N-V qualia ou non. Celle-ci ne pourrait en effet se faire simplement avec une technique d'apprentissage non relationnelle, notamment parce que le nombre d'attributs décrivant un couple est variable selon le nombre et la nature des mots apparaissant dans son contexte. Par ailleurs, l'expressivité de la PLI et la possibilité d'adjoindre des informations via le *background knowledge* permettent également de décrire et d'exploiter aisément les données relationnelles telles que nos structures hiérarchiques d'étiquettes catégorielles et sémantiques. Enfin, cette technique offre la possibilité de gérer des données bruitées, ce qui se révèle indispensable pour notre tâche où des erreurs, même en taux faibles, sont inhérentes à nos processus d'étiquetage.

Le logiciel de PLI utilisé lors de nos expérimentations est ALEPH, une implémentation en Prolog réalisée par Ashwin Srinivasan[4] [SRI 01]. Son fonctionnement peut être simplement décrit par l'algorithme 1 donné ci-dessous.

Algorithme 1. Algorithme d'ALEPH

Itération jusqu'à $E^+ = \emptyset$

1) choisir aléatoirement un exemple positif e^+ dans E^+ ;

2) définir un espace de recherche d'hypothèses $\mathscr{E}_H$ à partir de e^+ et du langage d'hypothèses $\mathscr{L}_H$;

3) parcourir l'espace de recherche $\mathscr{E}_H$ à la recherche de la clause h maximisant une fonction de score Sc ;

4) ajouter h à l'ensemble H et ôter de E^+ les exemples couverts par h.

Fin itération

4. ALEPH est disponible à l'URL http ://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/.

3.3. *Apprentissage de patrons d'extraction*

Comme nous l'avons expliqué précédemment, les algorithmes de PLI génèrent des règles expliquant ce qui caractérise les exemples du concept à apprendre par rapport aux contre-exemples. Dans notre cas, nous désirons discriminer les couples N-V qualia des non-qualia en fonction de leur contexte catégoriel et sémantique. Notre première tâche est donc de construire un jeu d'exemples positifs et négatifs décrivant les phrases dans lesquelles apparaissent des couples N-V qualia ou non-qualia en termes d'informations morphosyntaxiques et sémantiques. Nous explicitons dans un premier temps la constitution de ces exemples et contre-exemples, avant de présenter la façon dont l'algorithme de PLI infère des généralisations caractérisant les exemples positifs.

3.3.1. *Constitution et représentation des exemples*

Pour constituer les exemples, chaque occurrence en contexte d'un sous-ensemble de couples N-V de notre corpus est manuellement examinée par un expert. Les occurrences considérées comme qualia forment ainsi l'ensemble E^+, et les non-qualia l'ensemble E^- ; leur contexte (informations catégorielles et sémantiques de tous les mots des phrases dans lesquelles elles apparaissent) est décrit dans le *background knowledge* $\mathcal{B}$.

Par exemple, la phrase contenant le couple N-V qualia en gras « *L'installation se compose : de deux atterrisseurs* ***protégés*** *par des* ***carénages****, fixés et articulés...* » est transformée en

```
is_qualia(m11124_52,m11124_35).
```

qui est ajouté à E^+ (m11124_52 et m11124_35 étant respectivement les identificateurs du nom et du verbe) et les informations suivantes sont ajoutées à $\mathcal{B}$ (pour faciliter la compréhension, nous avons noté à droite les mots de la phrase décrits par les prédicats tags/3) :

```
sentence_beginning(m11123_3).
tags(m11123_3, tc_noun_sg, ts_pro).          installation
tags(m11123_16, tc_pron, ts_ppers).          se
pred(m11123_16, m11123_3).
tags(m11123_19, tc_verb_sg, ts_posv).        compose
pred(m11123_19, m11123_16).
tags(m11123_27, tc_wpunct_pf, ts_ponct).     :
pred(m11123_27, m11123_19).
tags(m11124_1, tc_prep, ts_rde).             de
pred(m11124_1, m11123_27).
tags(m11124_4, tc_num, ts_quant).            deux
pred(m11124_4, m11124_1).
tags(m11124_9, tc_noun_pl, ts_art).          atterriseurs
pred(m11124_9, m11124_4).
tags(m11124_35, tc_verb_adj, ts_acp).        protégés
```

pred(m11124_35, m11124_9).	
tags(m11124_44, tc_prep, ts_rman).	*par*
pred(m11124_44, m11124_35).	
tags(m11124_52, tc_noun_pl, ts_art).	***carénages***
pred(m11124_52, m11124_44).	
tags(m11124_62, tc_wpunct, ts_virg).	,
pred(m11124_62, m11124_52).	
tags(m11125_1, tc_verb_adj, ts_acp).	*fixés*
pred(m11125_1, m11124_62).	
tags(m11125_7, tc_conj_coord, ts_rconj).	*et*
pred(m11125_7, m11125_1).	
tags(m11125_10, tc_verb_adj, ts_acp).	*articulés*
pred(m11125_10, m11125_7).	
...	
distances(m11124_52, m11125_35, 2,1).	...

Dans cet exemple, pred(x,y) indique que le mot y apparaît juste avant le mot x dans la phrase, le prédicat tags/3 donne les étiquettes catégorielle et sémantique d'un mot, sentence_beginning/1 et sentence_end/1 marquent les mots apparaissant en début ou en fin de la phrase, et distances/4 spécifie le nombre de mots et de verbes entre N et V (une distance négative indique que N apparaît avant V, une valeur positive que V apparaît avant N ; ces valeurs sont décalées d'une unité pour refléter l'ordre entre N et V lorsque les distances sont nulles). On constate que certaines catégories de mots ne sont pas prises en compte. C'est notamment le cas des déterminants et de certains adjectifs, qui ne sont pas considérés pertinents pour caractériser la nature qualia d'une paire N-V au sein d'une phrase. Les exemples négatifs sont codés de la même façon. Près de 3 100 exemples positifs et 3 200 négatifs sont ainsi produits à partir du corpus MATRA-CCR.

Par ailleurs, les informations relatives à la hiérarchie des classes catégorielles et sémantiques sont transcrites dans $\mathcal{B}$. Par exemple, le fait qu'un mot ayant une étiquette tc_verb_pl soit un verbe conjugué au pluriel et puisse être considéré comme un verbe conjugué, et plus généralement comme un verbe, est précisé en Prolog par :

```
conjugated_plural_verb(W) :- tags(W, tc_verb_pl, _).

conjugated_verb(W) :- conjugated_plural_verb(W).

verb(W) :- conjugated_verb(W).
```

Ce sont ces informations, les littéraux conjugated_plural_verb(W), conjugated_verb(W), et bien d'autres, qui vont former les éléments constitutifs de nos patrons, les variables représentant des mots (voir les exemples de patrons en section suivante).

3.3.2. *Parcours de l'espace de recherche*

Comme nous l'avons dit, la plupart des algorithmes de PLI permettent de spécifier un langage d'hypothèses pour définir la forme attendue des règles inférées. Dans notre cas, ce langage d'hypothèses exploite les informations catégorielles et sémantiques des mots apparaissant dans les exemples (c'est-à-dire N, V ou leur contexte) ainsi que les informations d'ordre et de distances entre ces mots. Une description détaillée de ce langage, et des techniques utilisées pour contrôler l'expressivité de notre algorithme d'inférence de patrons, est donnée dans [CLA 03b]. Il permet principalement de s'assurer de la pertinence linguistique des patrons proposés, en évitant par exemple que des informations redondantes sur un mot soient données (contrainte de concision) ou, qu'au contraire, un mot soit introduit dans le patron sans qu'aucune information lexicale, morpho-syntaxique ou sémantique ne lui soit associée (contrainte d'utilisation des variables).

Comme indiqué au paragraphe 3.2, l'ensemble des hypothèses décrivant un exemple et répondant aux contraintes du langage d'hypothèses peut être organisé par une relation de généralité entre hypothèses (on parle alors de subsomption). Celle-ci permet de structurer l'espace de recherche en vue d'un parcours efficace. En effet, la notion de subsomption que nous avons définie pour ce faire est un quasi-ordre qui, associé à la forme des hypothèses manipulées, implique que $\mathscr{E}_H$ est un treillis (voir la démonstration dans [CLA 03b]). De manière simplifiée, cette subsomption permet de modéliser le fait qu'une hypothèse C est plus « générale » qu'une autre D si :

1) soit C est constituée d'une sous-partie des littéraux présents dans D ;

2) soit C contient des littéraux plus généraux que ceux présents dans D.

La figure 2 présente un tel treillis d'hypothèses (simplifié à des fins de lisibilité). Les numéros apparaissant sur les arcs entre les hypothèses indiquent à quelle notion de subsomption présentée ci-dessus se rapporte le lien de généralité entre deux clauses.

L'espace de recherche $\mathscr{E}_H$ défini par notre langage d'hypothèses demeure encore trop vaste pour être exploré extensivement. Dans [CLA 03a], nous présentons l'opérateur de raffinement que nous avons mis au point pour parcourir notre espace de recherche de manière performante en tirant le meilleur parti de sa structure de treillis.

La fonction de score Sc que nous avons retenue, permettant de trouver la meilleure hypothèse au sein de $\mathscr{E}_H$ (*cf.* algorithme 1, étape 3), dépend de $|E_h^+|$ et $|E_h^-|$, mais également d'autres paramètres comme la longueur L de l'hypothèse h testée. Plus précisément, elle est définie par le couple $\left(|E_h^+| - |E_h^-|, L\right)$; une hypothèse h_1 de score $\left(|E_{h_1}^+| - |E_{h_1}^-|, L_1\right)$ est dite meilleure qu'une hypothèse h_2 de score $\left(|E_{h_2}^+| - |E_{h_2}^-|, L_2\right)$ si et seulement si $|E_{h_1}^+| - |E_{h_1}^-| > |E_{h_2}^+| - |E_{h_2}^-|$ ou $|E_{h_1}^+| - |E_{h_1}^-| = |E_{h_2}^+| - |E_{h_2}^-| \wedge L_1 < L_2$.

Enfin, comme pour toute technique d'apprentissage artificiel, il est important de contrôler le niveau de généralisation des règles obtenues. Une trop grande généralisation engendrerait une faible précision de l'extraction, et une faible généralisation (apprentissage par cœur) un faible rappel. Ce contrôle et le réglage des paramètres qui

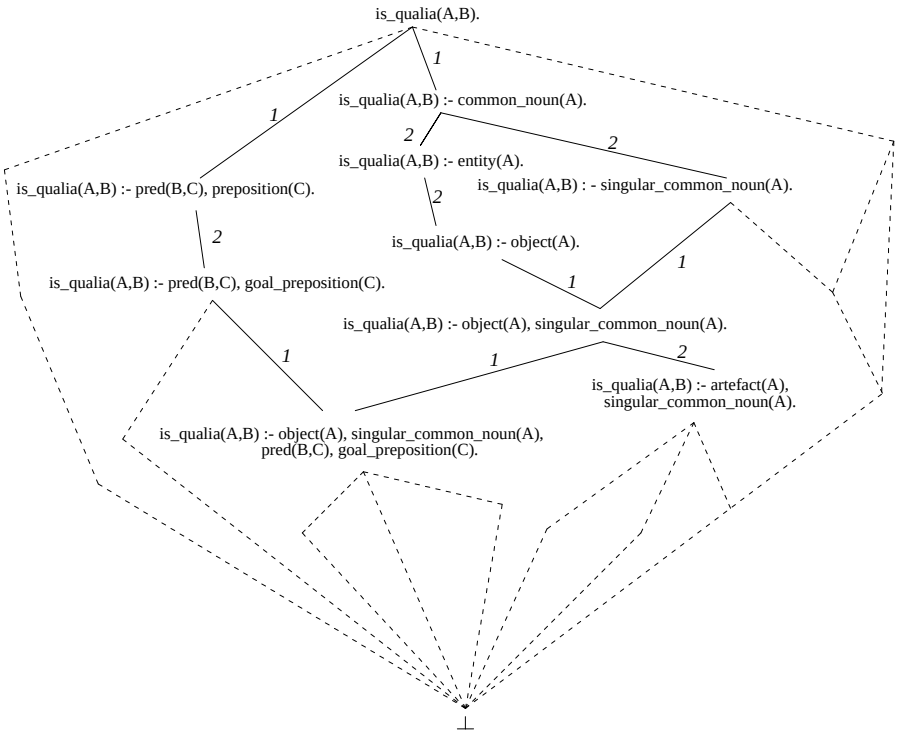

Figure 2. *Organisation en treillis de l'espace de recherche $\mathscr{E}_H$*

en découle — notamment le bruit autorisé — se font de manière automatique grâce à une validation croisée en 10 blocs [KOH 95] : l'ensemble des exemples et contre-exemples d'apprentissage (E^+ et E^-) est divisé aléatoirement en 10 sous-ensembles ; chaque sous-ensemble sert alternativement de jeu de test pour évaluer l'apprentissage effectué à partir des 9 autres sous-ensembles. Les résultats de chaque évaluation sont résumés dans une matrice de confusion similaire à celle de la figure 1[5]. On peut ainsi calculer le coefficient Φ qui traduit en une seule grandeur toutes les informations de cette matrice de confusion :

$$\Phi = \frac{(TP * TN) - (FP * FN)}{\sqrt{PrP * PrN * AP * AN}}$$

Une moyenne de Φ est calculée à partir des 10 matrices obtenues sur les 10 sous-ensembles de test. Pour chaque valeur possible des paramètres de l'algorithme de

5. La signification des variables est donnée par la combinaison des lettres : A signifie *actual* (réel), Pr *predicated* (prédit), T *true* (vrai), F *false* (faux), P *positive* (positif) et N *negative* (négatif) ; S est le total.

	qualia réel	non-qualia réel	Total
prédit qualia	TP	FP	PrP
prédit non-qualia	FN	TN	PrN
Total	AP	AN	S

Tableau 1. *Matrice de confusion*

PLI, une validation croisée est donc effectuée et un Φ moyen calculé. Les valeurs des paramètres retenues sont celles qui maximisent ce Φ moyen. Le tableau 2 présente les résultats obtenus[6] avec cette valeur optimale.

	Temps (secondes)	Précision (%)	Rappel (%)	Coefficient Φ
Moyenne	2 198	81,09	88,81	0,70
Écart-type	830	2,34	1,74	0,01

Tableau 2. *Résultats de la validation croisée*

L'emploi de la validation croisée apporte donc deux précieuses indications. Elle permet, d'une part, d'évaluer la qualité intrinsèque de l'apprentissage effectué et, d'autre part, de régler automatiquement certains des paramètres nécessaires à l'algorithme de PLI.

L'apprentissage est alors relancé avec ces réglages et la totalité des exemples. Finalement, les règles produites, ensuite utilisées comme patrons pour extraire de nouvelles paires qualia, sont par exemple de la forme suivante :

```
is_qualia(N,V) :- infinitive(V), action_verb(V), artefact(N), pred(V,A), pred(A,N).
```

Cette règle signifie qu'une paire composée d'un nom N et d'un verbe V est considérée qualia si N est un artefact apparaissant dans la phrase avant un mot A quelconque, lui-même précédant le verbe d'action à l'infinitif V. Une telle règle — dont la succession des pred/2 traduit l'aspect relationnel difficilement capturable par une méthode d'apprentissage de type attribut-valeur — permet ainsi d'extraire le couple qualia *prise-alimenter* dans la phrase *« La prise peut alimenter une pompe à carburant... »* et le couple *vis-serrer* dans *« ...écrous, vis : serrer au couple... »*

6. La machine utilisée lors de ces mesures est un PC Pentium IV 2,4 GHz, 512 Mo de RAM sous Red Hat Linux 8.0.

3.4. *Évaluation des performances d'extraction*

Cette section présente les performances de notre système symbolique ASARES en conditions réelles d'acquisition de ressources linguistiques grâce à un jeu de test que nous décrivons ci-dessous. La validité linguistique et l'expressivité des clauses produites sont discutées en dernière partie.

3.4.1. *Construction du jeu de test*

Le jeu de test sur lequel nous évaluons les performances de notre système d'acquisition de couples qualia est un extrait de 32 000 mots du corpus MATRA-CCR étiqueté catégoriellement et sémantiquement. Malgré sa taille relativement petite, examiner manuellement toutes les occurrences de couples N-V de ce sous-corpus pour les annoter comme qualia ou non-qualia est impossible. Nous nous sommes donc concentrés sur 7 noms particulièrement représentatifs du vocabulaire du corpus : *vis, écrou, porte, voyant, prise, capot, bouchon.* Ces noms sont tous parmi les termes jugés les plus spécifiques du corpus par le logiciel d'extraction de termes TERMOSTAT [DRO 03]. Leurs classes sémantiques sont par ailleurs représentatives de celles des termes les plus spécifiques de ce corpus, à savoir principalement `artefact`, `instrument`, `objet physique`. Pour ne pas fausser les mesures, aucun de ces noms n'a évidemment été utilisé lors des phases d'apprentissage.

Un programme Perl présente toutes les occurrences de couples N-V, où N est l'un des 7 noms recherchés, apparaissant au sein d'une phrase du sous-corpus, à quatre experts qui annotent alors ces paires comme qualia ou non-qualia. Les divergences sont discutées jusqu'à ce qu'un accord se dégage. Finalement, parmi les 286 couples différents trouvés, 66 d'entre eux sont notés comme étant qualia. Ce jeu de test est ensuite utilisé pour comparer les résultats d'extraction de notre système à ceux des experts.

3.4.2. *Évaluation des résultats empiriques*

La comparaison entre le jeu de test et les couples retenus par application des patrons obtenus grâce à ASARES sur le sous-corpus se fait à l'aide de matrices de confusion telles que celle donnée précédemment. On peut ainsi, grâce à ces matrices, calculer les taux de rappel et de précision. Cependant, il faut préalablement choisir un seuil (noté s par la suite), c'est-à-dire un nombre minimum d'occurrences détectées par nos patrons appris, à partir duquel un couple extrait sera effectivement considéré comme qualia. Ainsi, les taux de rappel R et de précision P du système d'acquisition, calculés grâce à notre jeu de test, s'expriment en fonction de s (mêmes notations que précédemment) par :

$$R(s) = \frac{TP(s)}{TP(s) + FN(s)},\ P(s) = \frac{TP(s)}{TP(s) + FP(s)}.$$

Un seuil bas aura évidemment tendance à favoriser le rappel au détriment de la précision et un seuil élevé produira l'effet l'inverse. Pour représenter les performances

en fonction des différentes valeurs de s possibles (on note $\mathcal{S}$ cet ensemble), on utilise usuellement les courbes rappel-précision dans lesquelles chaque point représente la précision du système étant donné son rappel pour un seuil s donné.

La figure 3 représente la courbe rappel-précision obtenue par ASARES sur le jeu de test décrit précédemment. La référence utilisée comme point de comparaison (*baseline*) est la densité de couples qualia parmi tous les couples N-V étudiés du sous-corpus, c'est-à-dire $\frac{AP}{S} = \frac{66}{286} = 0,231$. Cette densité représente la précision moyenne qu'obtiendrait un système choisissant au hasard les couples qualia.

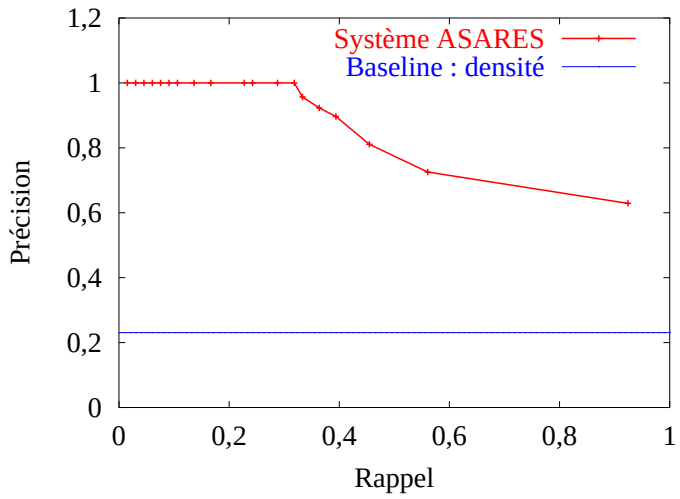

Figure 3. *Courbes rappel-précision du système* ASARES

Pour faciliter les comparaisons entre systèmes, on cherche parfois à leur assigner une mesure unique. Nous utilisons dans ce but la mesure Φ définie précédemment, qui synthétise en une seule grandeur les caractéristiques du système pour un seuil donné. Nous calculons également la F-mesure, fréquemment utilisée dans le domaine de la recherche d'information. Elle est définie comme étant la moyenne harmonique pondérée du taux de rappel ($R(s)$) et du taux de précision ($P(s)$). Dans le cas où un poids égal est donné au rappel et à la précision, la F-mesure s'écrit : $F(s) = \frac{2P(s)R(s)}{P(s)+R(s)}$. Le tableau 3 présente les taux de rappel, de précision, la F-mesure et le coefficient Φ de notre système d'acquisition symbolique de couples qualia. Le seuil s_{opt} retenu est choisi tel que $\Phi(s_{opt})$ soit maximal ($s_{opt} = \underset{s \in \mathcal{S}}{\text{argmax}}(\Phi(s))$). Dans notre cas, s_{opt} est égal à 1 ; on obtient les meilleures performances en considérant un couple qualia dès qu'il est détecté au moins une fois.

	rappel (%)	précision (%)	F-mesure	coefficient Φ
PLI supervisé	92,4	62,2	0,744	0,671

Tableau 3. *Performances du système* ASARES *supervisé*

3.4.3. *Évaluation linguistique*

Comme nous l'avons dit précédemment, le double but de nos travaux est de définir une technique performante d'extraction de couples N-V qualia, mais aussi de fournir un support linguistique au concept de rôle qualia. La validation de notre système symbolique présenté ci-avant passe donc également par une évaluation de l'intérêt linguistique des règles générées[7].

L'apprentissage mené par ASARES sur les exemples positifs et négatifs fournis en entrée conduit à l'ensemble de neuf clauses suivant, que nous interprétons dans une perspective linguistique :

1) is_qualia(A,B) :- precedes(B,A), near_verb(A,B), infinitive(B), action_verb(B).

2) is_qualia(A,B) :- contiguous(A,B).

3) is_qualia(A,B) :- precedes(B,A), near_word(A,B), near_ verb(A,B), suc(B,C), preposition(C).

4) is_qualia(A,B) :- near_word(A,B), sentence_beginning(A).

5) is_qualia(A,B) :- precedes(B,A), singular_common_noun(A), suc(B,C), colon(C), pred(A,D), punctuation(D).

6) is_qualia(A,B) :- near_word(A,B), suc(B,C), suc(C,D), action_verb(D).

7) is_qualia(A,B) :- precedes(A,B), near_word(A,B), pred(A,C), punctuation(C).

8) is_qualia(A,B) :- near_verb(A,B), pred(B,C), pred(C,D), pred(D,E), preposition(E), sentence_beginning(A).

9) is_qualia(A,B) :- precedes(A,B), near_verb(A,B), pred(A,C), subordinating_conjunction(C).

Dans ces règles, le prédicat precedes(X,Y) signifie X apparaît dans la phrase avant Y ; pred(X,Y) indique que Y est le mot précédant immédiatement X et inversement suc(Y,X) précise que X suit immédiatement Y dans la phrase ; near_word(X,Y) montre que X et Y doivent être séparés par au moins un mot et au plus deux mots, et near_verb(X,Y) qu'il n'y a aucun verbe entre X et Y. Les variables A et B représentent respectivement le nom N et le verbe V étant en relation qualia. Le premier patron correspond par exemple au schéma : *V d'action à l'infinitif + (tout sauf un verbe)* + N*.

On constate que les règles inférées font ressortir des schémas très généraux de proximité entre les constituants des couples (au plus un mot doit séparer le N du V) ou

7. P. Bouillon (ETI/TIM/ISSCO, Genève) et C. Fabre (ERSS, Toulouse) ont largement participé à cette évaluation linguistique.

de position (V doit apparaître avant N dans la phrase). Par ailleurs, peu d'informations sémantiques sont utilisées à ce niveau de généralisation, à l'exception notable des verbes (les verbes d'action sont privilégiés). D'autres informations surfaciques généralement délaissées des linguistes, comme les ponctuations particulièrement présentes et structurantes dans notre texte très technique, sont également exploitées dans des schémas plus spécifiques à notre corpus. Ainsi, la clause 5 met en évidence les structures de listes très nombreuses dans le corpus et couvre par exemple les phrases de la forme « *...Verbe : ... , Nom au singulier...* ». L'emploi de nombreux verbes à l'infinitif, typique des instructions composant le corpus, est également souligné dans la plupart des règles générées. Les clauses obtenues présentent donc, d'un point de vue linguistique, un certain nombre de caractéristiques très génériques mais aussi, comme nous nous y attendions, certaines spécificités propres à notre corpus. Une description plus détaillée des patrons inférés par cette approche supervisée sur le corpus MATRA CCR et des phénomènes linguistiques qu'elle permet effectivement de prendre en compte est donnée dans [BOU 02a].

3.5. *Bilan*

Le système ASARES produit donc des résultats satisfaisants, à la fois en termes de qualité des éléments acquis sur corpus et d'interprétabilité des patrons inférés grâce à l'utilisation de la PLI. Il répond ainsi, comme nous l'avons fait remarquer, à deux de nos trois critères initiaux. L'approche d'apprentissage adoptée autorise par ailleurs une souplesse d'utilisation permettant d'appliquer cette technique à l'extraction d'autres éléments sémantiques. Cela permet une certaine généricité de notre outil mais ne suffit pas à satisfaire complètement notre dernier critère : la portabilité et l'automaticité, mises à mal par la nécessité de fournir manuellement en entrée de notre technique d'acquisition de nombreux exemples et contre-exemples de couples qualia. La section suivante présente deux solutions à ce problème inhérent à la PLI, obtenues en combinant notre méthode symbolique à une méthode statistique d'acquisition.

4. Apprentissage semi-supervisé

De nombreux travaux actuels visent à réduire les coûts dus à la constitution des exemples dans les méthodes d'apprentissage supervisé. La plupart de ces travaux s'appuient sur des variantes de *bootstrapping* [JON 99], et les techniques d'apprentissage résultantes sont alors dites semi-supervisées. À partir de principes similaires, nous proposons deux variantes de notre système d'extraction, travaillant sur le même corpus étiqueté catégoriellement et sémantiquement, mais qui permettent, grâce à l'adjonction d'une technique classique d'extraction statistique que nous présentons ci-dessous, de supprimer la phase manuelle de construction des exemples et contre-exemples de couples qualia. Le paragraphe 4.2 explique successivement le fonctionnement de nos deux variantes semi-supervisées. Nous terminons en comparant leurs performances avec celles d'ASARES.

4.1. *Extraction statistique*

De nombreux travaux d'acquisition d'informations à partir de textes, et plus particulièrement d'extraction de cooccurrences, ont été menés, comme nous l'avons évoqué en section 2, via des approches statistiques [MAN 99, PEA 02]. Notre problème d'extraction peut s'inscrire dans ce cadre, les couples N-V qualia étant alors vus comme un type spécial de cooccurrences. Les expériences rapportées dans [BOU 02a] présentent les résultats obtenus pour l'extraction de paires qualia par quelques indices statistiques parmi les plus communément employés pour ce type de tâche (Dice, Kulczinsky, Ochiai, Yule, Loglike, *Simple Matching*, Information Mutuelle, Information Mutuelle au cube, Φ^2, Jaccard, cosinus, McConnoughy). Parmi ceux-ci, le coefficient d'Information Mutuelle au cube (IM^3 par la suite) proposé par B. Daille [DAI 94], dont nous rappelons la définition ci-après, donne les meilleurs résultats pour notre tâche (voir [BOU 02a] pour le détail des évaluations pour chacun de ces coefficients). Étant donné la table de contingence 4 (les cooccurrences sont calculées dans une fenêtre d'une phrase à partir des lemmes des mots), le coefficient IM^3 du couple N_i-V_j est donné par : $log_2 \frac{a^3}{(a+b)(a+c)}$.

	V_j	V_k , $k \neq j$
N_i	a	b
N_l , $l \neq i$	c	d

Tableau 4. *Table de contingence de la paire nom-verbe N_i-V_j*

Les résultats d'extraction de couples qualia avec cette technique [BOU 02a] se révèlent certes moins bons que l'approche PLI supervisée (jusqu'à 58 % de précision en moins pour un rappel fixé), et ce type de méthodes d'extraction ne fournit aucun élément de compréhension sur les résultats produits et ne peut donc répondre directement à notre problématique. Cependant, cette approche statistique possède des avantages intéressants : elle est totalement automatique (aucune intervention humaine n'est requise), facile d'utilisation et donc tout à fait portable d'un corpus à un autre. La partie suivante expose deux techniques pour transposer ces avantages vers notre système symbolique d'extraction.

4.2. *Approches mixtes*

Pour remplir notre double tâche de construction d'éléments du Lexique génératif et de constitution de patrons d'extraction linguistiquement motivés, nous proposons donc de combiner les avantages des deux méthodes précédentes en deux systèmes d'extraction mixtes. Ces systèmes conservent notamment l'aspect non supervisé et donc entièrement automatique du cadre statistique tout en gardant l'aspect explicatif du cadre symbolique grâce à la production de règles pertinentes. La sous-section suivante présente une première combinaison d'ASARES avec une méthode d'extraction statistique reposant sur un échange séquentiel des résultats entre chacune des

méthodes. Le paragraphe 4.2.2 expose quant à lui un second moyen de combiner les deux techniques d'acquisition en incluant plus directement les résultats de l'approche statistique au sein de la phase d'inférence des patrons d'extraction d'ASARES.

4.2.1. *Approche mixte séquentielle*

La technique d'extraction mixte présentée dans cette partie repose sur une combinaison séquentielle des systèmes symbolique et statistique présentés précédemment. Comme il est indiqué dans l'algorithme 2, chaque système utilise itérativement en entrée les données de sortie de l'autre système. Plus précisément, la liste de paires N-V générée par un système (L_{PLI} pour le symbolique, L_{IM^3} pour le statistique) est utilisée par l'autre pour construire sa propre liste de couples. La seule contrainte est de débuter cette itération avec la méthode statistique puisqu'elle ne nécessite aucune donnée autre que le corpus. À l'initialisation, tous les couples N-V apparaissant au sein d'une phrase sont considérés comme potentiellement qualia ; cela est indiqué grâce à la règle is_qualia(N,V). donnée dans la liste de patrons d'extraction L_R.

Algorithme 2. Système mixte séquentiel

Initialisation

· L_R = {is_qualia(N,V).}

· application des règles de L_R au corpus ; les couples N-V extraits et leur nombre d'occurrences détectées sont insérés dans L_{PLI}

Itération

1) pour tout couple $N_i - V_j$ de L_{PLI}

· construction de la table de contingence de $N_i - V_j$ avec les nombres d'occurrences indiqués dans L_{PLI}

· calcul du score de $N_i - V_j$ selon IM^3

· insertion, suivant son score, du couple dans la liste triée décroissante L_{IM^3}

2) constitution de l'ensemble E^+ (respectivement E^-) à partir de toutes les occurrences dans le corpus des n_1 (resp. n_2) premiers (resp. derniers) couples de L_{IM^3}

3) apprentissage par PLI avec E^+ et E^- ; les règles obtenues sont regroupées dans L_R

4) application des règles de L_R au corpus, les couples N-V extraits et leur nombre d'occurrences détectées sont réunis dans L_{PLI}

L'itération s'arrête lorsque le même ensemble de règles est obtenu lors de deux tours successifs. Lors de nos expériences, n_1 a été choisi (à chaque itération) tel que les n_1 premiers couples de L_{IM^3} soient tous ceux ayant un score d'association positif ; n_2 a quant à lui été choisi tel que $n_2 = n_1$. Le système d'extraction résultant est appelé par la suite système mixte séquentiel.

4.2.2. *Approche mixte intégrée*

Contrairement au système présenté ci-dessus dans lequel les techniques statistique et symbolique sont utilisées sans modifications majeures, le second système mixte que nous proposons combine ces deux approches plus étroitement et nécessite quelques changement dans l'algorithme de PLI.

Comme nous l'avons mentionné au paragraphe 3.2, lors de la troisième étape d'un apprentissage par PLI, une règle h est choisie parmi un espace d'hypothèses $\mathscr{E}_H$ si elle maximise une fonction de score Sc. Cette fonction dépend du nombre d'exemples positifs et négatifs que h couvre ; ainsi, on a :

$$h = \underset{h \in \mathscr{E}_H}{\text{argmax}} \; Sc\left(|E_h^+|, |E_h^-|\right).$$

Le principe de notre seconde méthode mixte est de pondérer les exemples selon leur score statistique. Les hypothèses sont donc désormais évaluées à partir des poids des exemples (que nous définissons ci-dessous) qu'elles couvrent. Les ensembles d'exemples et contre-exemples sont issus des résultats d'extraction de la méthode d'extraction IM^3 : toutes les occurrences dans le corpus des couples ayant les plus hauts scores sont codées dans E^+, et inversement, celles ayant les scores les plus faibles sont placées dans E^- ; un poids w, calculé à partir des scores IM^3, est assigné à chacun de ces exemples. Plus précisément, le poids d'un exemple est le score IM^3 du couple N-V qu'il contient, normalisé de telle manière que la somme des poids des exemples positifs soit égale à la somme des poids des exemples négatifs, soit :

$$\sum_{e^+ \in E_h^+} w(e^+) = \sum_{e^- \in E_h^-} w(e^-).$$

Ainsi, plus un exemple est considéré comme pertinent (c'est-à-dire ayant un score important) par la méthode statistique, plus il influencera le choix des hypothèses. Finalement, les règles choisies sont celles maximisant $Sc(h)$ redéfinie par :

$$h = \underset{h \in \mathscr{E}_H}{\text{argmax}} \, Sc\left(\sum_{e^+ \in E_h^+} w(e^+), \sum_{e^- \in E_h^-} w(e^-)\right)$$

Avec ces paramétrages et les ensembles E^+ et E^- générés automatiquement, l'algorithme de PLI modifié se déroule comme indiqué en 3.2 et produit ainsi des règles utilisées ensuite comme patrons d'extraction. Cette technique est appelée par la suite système mixte intégré.

4.3. *Évaluation des performances*

Nous évaluons comme précédemment les résultats obtenus par nos deux techniques semi-supervisées. Nous examinons donc dans un premier temps les perfor-

mances d'extraction qu'elles obtiennent sur notre jeu de test puis, dans un second temps, les schémas linguistiques portés par les patrons qu'elles infèrent.

4.3.1. *Performances d'extraction des systèmes mixtes*

La figure 4 présente les courbes rappel-précision pour nos deux systèmes d'extraction symbolique semi-supervisés ; les systèmes ASARES supervisé et IM^3 servent de référence. Pour ce dernier système, comme pour toute méthode d'acquisition statistique, il est nécessaire comme précédemment de définir une valeur-seuil s à partir de laquelle un couple est considéré comme qualia ; les paires dont le score statistique est inférieur à s sont alors considérées non qualia. Les valeurs de la courbe sont obtenues avec s optimal.

On remarque que les performances de nos systèmes symboliques semi-supervisés sont très proches de la version supervisée et donc nettement supérieures à celles du système statistique, notamment lorsque le rappel est élevé. Plus précisément, on constate que la version semi-supervisée mixte intégrée obtient une meilleure précision pour de faibles rappels alors qu'à l'inverse la technique mixte séquentielle affiche une précision sensiblement supérieure pour des rappels élevés.

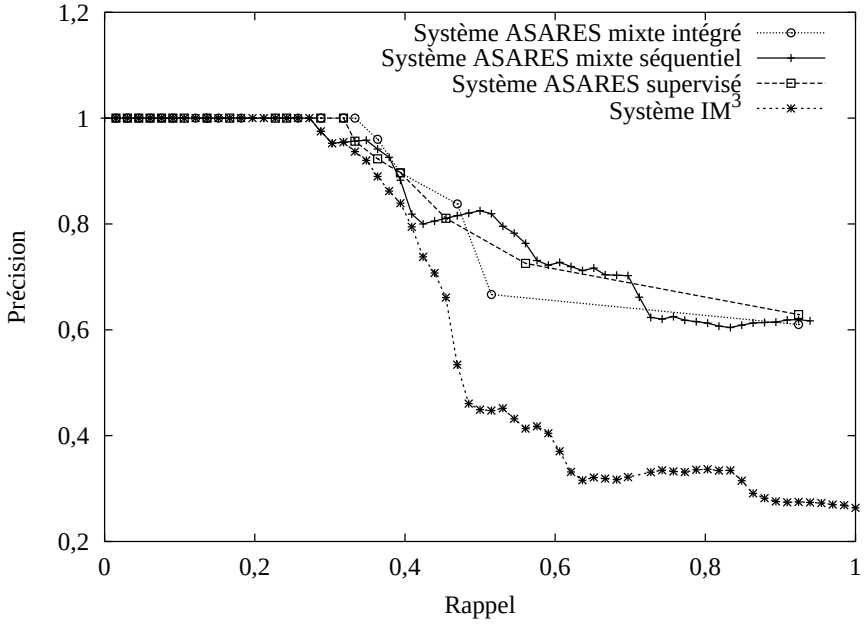

Figure 4. *Courbes rappel-précision des systèmes IM^3,* ASARES *supervisé, mixte séquentiel et mixte intégré*

Le tableau 5 compare les taux de rappel, de précision, la F-mesure et le coefficient Φ de nos systèmes mixtes avec ceux des systèmes ASARES supervisé et IM^3. Le

seuil s_{opt} retenu pour chacun des systèmes est choisi tel que $\Phi(s_{opt})$ soit maximal ($s_{opt} = \underset{s \in \mathcal{S}}{\text{argmax}}(\Phi(s))$).

	rappel (%)	précision (%)	F-mesure	coefficient Φ
ASARES supervisé	92,4	62,2	0,744	0,671
IM^3	36,4	92,3	0,522	0,520
mixte séquentiel	93,9	62,0	0,747	0,677
mixte intégré	89,4	60,2	0,720	0,636

Tableau 5. *Performances des méthodes* ASARES *supervisé,* IM^3 *et mixtes*

Enfin, à titre informatif, la figure 5 représente l'évolution des performances du système mixte séquentiel au cours de sa construction, c'est-à-dire lors du déroulement de l'algorithme 2 sur le jeu de test décrit précédemment. Plus précisément, après chaque étape 1 et 4 de l'algorithme, on recherche comme précédemment le seuil s_{opt} maximisant Φ à partir des listes de couples qualia L_{IM^3} et L_{PLI}, et on calcule également la F-mesure pour ce même seuil. On constate qu'il suffit de trois itérations pour approcher les performances du systèmes d'extraction PLI supervisé ; la convergence empirique de l'algorithme est donc très rapide.

4.3.2. *Évaluation linguistique*

La validation des systèmes symboliques présentés ci-avant passe comme précédemment par une évaluation de l'intérêt linguistique des règles générées. À ce titre, on note tout d'abord de très grandes similarités entre les règles produites par nos deux systèmes semi-supervisés et la version originale d'ASARES. Cela explique bien entendu la similitude, constatée ci-dessus, de leurs performances pour la tâche d'extraction. On retrouve donc dans ces règles des schémas très généraux de proximité entre les constituants des couples, de position, ainsi que l'exploitation des indices surfaciques comme les ponctuations dans des schémas plus spécifiques à notre corpus. Peu d'informations sémantiques sont également utilisées à ce niveau de généralisation, à l'exception notable des verbes puisque comme précédemment les verbes d'action sont privilégiés.

Les deux systèmes semi-supervisés répondent donc à nos attentes. Ils combinent en cela les avantages des deux approches sur lesquelles ils reposent : l'approche statistique permettant l'automatisation du processus et l'approche symbolique garantissant une bonne qualité des résultats et des règles d'extraction produites.

5. Discussion et perspectives

Nous avons présenté une approche symbolique pour l'acquisition sur corpus de ressources linguistiques qui présente l'avantage de fournir un classifieur, efficace pour

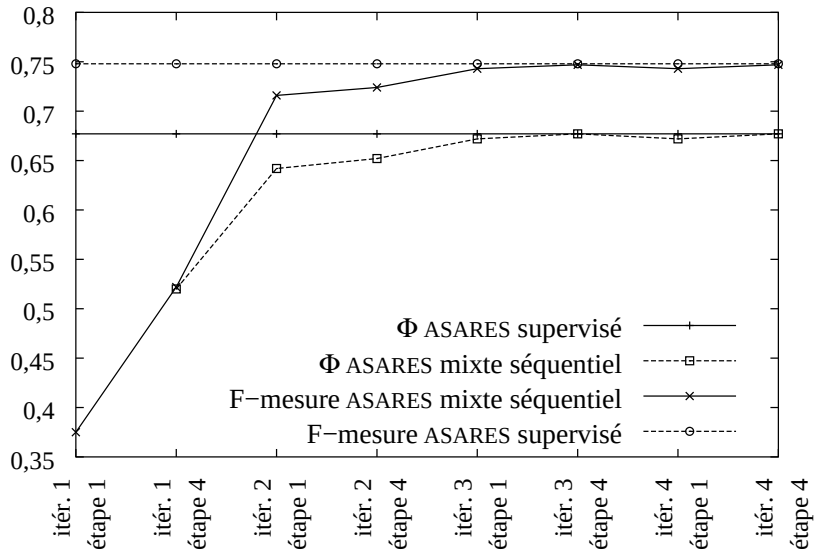

Figure 5. *Évolution des performances du système mixte séquentiel au cours de l'algorithme*

la tâche d'extraction de couples N-V qualia, et également interprétable, contrairement aux méthodes statistiques communément employées pour ce type de tâche. L'utilisation de la PLI comme technique d'apprentissage présente le double avantage de bénéficier d'une grande expressivité à la fois pour définir notre problème et pour la génération de patrons d'extraction linguistiquement fondés. Comme nous l'avons montré avec nos deux systèmes mixtes, l'obstacle majeur à l'utilisation de cette technique d'apprentissage supervisé, à savoir la phase manuelle de construction des exemples, peut être contourné par l'emploi d'une technique d'extraction statistique servant à amorcer le système symbolique, les systèmes mixtes obtenus conduisant à des performances similaires à celles de la version supervisée.

Ces systèmes semi-supervisés n'échappent cependant pas aux coûts des étiquetages, notamment sémantiques, du corpus. Néanmoins, comme le suggère les expériences rapportées dans [CLA 01] et la faible utilisation des informations sémantiques dans les règles produites, ce coût peut être contrôlé (en n'effectuant qu'un étiquetage sémantique partiel, sur quelques catégories de mots, par exemple) sans modifier profondément les résultats d'extraction. D'autres expériences d'acquisition sont actuellement en cours pour vérifier, dans un cadre linguistique différent et sur d'autres types de corpus, si ce type d'étiquetage sémantique apporte finalement un gain suffisant à la tâche d'acquisition par ASARES pour justifier sa mise en œuvre. Une autre

limite d'utilisation de ce type de système d'acquisition réside dans la taille requise des corpus à traiter. Si, par essence, la méthode symbolique fonctionne quel que soit le volume de texte (pourvu qu'il soit homogène dans ses structures sémantiques et morphosyntaxiques), l'utilisation d'un *bootstrap* statistique impose une taille minimale au corpus pour obtenir des résultats de cooccurrences fiables.

D'un point de vue plus théorique, les deux systèmes semi-supervisés que nous avons présentés se rapprochent de certaines versions évoluées de *bootstrapping* telles que le *co-training* [BLU 98] ou celle proposée par Yarowsky [YAR 95], sans en partager cependant les propriétés formelles. Ces deux dernières techniques assurent en effet des résultats théoriques intéressants d'apprenabilité, mais au prix de conditions contraignantes sur les données. Par exemple, le *co-training* impose que les données d'apprentissage puissent être représentées selon deux *vues* conditionnellement indépendantes. Malheureusement, cette forte condition d'indépendance est rarement avérée dans les données réelles [ABN 02]. Nos deux techniques semi-supervisées reposent néanmoins implicitement sur une condition analogue : pour éviter que la phase d'apprentissage par PLI ne soit biaisée, nous supposons que les différentes occurrences des couples apparaissent dans des structures sémantiques et morphosyntaxiques variées qui donneront naissance à nos patrons d'extraction. Or, notre corpus comporte de nombreuses instructions répétées à l'identique ; cette hypothèse d'indépendance entre les couples extraits statistiquement et les patrons générés est donc en partie invalidée. Cependant, la ressemblance entre les règles générées par le système supervisé et les deux versions semi-supervisées semble montrer une bonne tolérance de nos algorithmes à ce propos.

De nombreuses perspectives sont ouvertes sur ces travaux. La variabilité des patrons produits et de la qualité des résultats d'extraction selon les domaines et les genres des textes traités doit être analysée. Pour ce faire, nous avons débuté des expériences similaires sur un corpus plus généraliste composés d'articles de journaux comptant plus de 6 millions de mots. D'un point de vue applicatif, ces systèmes peuvent être aisément adaptés à l'acquisition d'autres types d'éléments textuels (termes complexes, collocations et autres informations lexicales). Dans une application à la structuration de terminologies, le système ASARES, version supervisée, a notamment été utilisé récemment avec de très bons résultats pour une tâche d'acquisition de certaines relations sémantiques exprimées par des fonctions lexicales [CLA 04a]. Enfin, sur un aspect plus théorique, l'utilisation de statistiques (vues comme une sorte de distribution de probabilités sur les ensembles d'exemples E^+ et E^-) en PLI, comme cela est fait dans le système mixte intégré, soulève d'intéressantes problématiques [MUG 94a].

6. Bibliographie

[ABN 02] ABNEY S., « Bootstrapping », *40th Annual Meeting of the Association for Computational Linguistics, ACL*, Philadelphia, PA, États-Unis, 2002.

[AGA 95] AGARWAL R., « Semantic Feature Extraction from Technical Texts with Limited Human Intervention », Thèse de doctorat, Mississippi State University, États-Unis, 1995.

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

[BLU 98] BLUM A., MITCHELL T., « Combining Labeled and Unlabeled Data with Co-training », *COLT : Proceedings of the Workshop on Computational Learning Theory*, Madison, WI, États-Unis, 1998.

[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, IC 97*, Roscoff, France, 1997.

[BOU 00a] BOUILLON P., BAUD R. H., ROBERT G., RUCH P., « Indexing by Statistical Tagging », *Journées d'Analyse statistique des Données Textuelles, JADT2000*, Lausanne, Suisse, 2000.

[BOU 00b] 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., Eds., *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.

[BOU 02a] BOUILLON P., CLAVEAU V., FABRE C., SÉBILLOT P., « Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method », *3rd International Conference on Language Resources and Evaluation, LREC 02*, Las Palmas de Gran Canaria, Espagne, 2002.

[BOU 02b] BOURIGAULT D., « Analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus », *Traitement automatique des langues naturelles, TALN'02*, Nancy, France, 2002.

[BRI 97] BRISCOE T., CARROLL J., « Automatic Extraction of Subcategorisation from Corpora », *5th ACL Conference on Applied Natural Language Processing, ANLP97*, Washington, États-Unis, 1997.

[CLA 01] CLAVEAU V., SÉBILLOT P., BOUILLON P., FABRE C., « Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? », *TAL (traitement automatique des langues), numéro spécial Lexiques sémantiques*, vol. 42, n° 3, 2001.

[CLA 03a] CLAVEAU V., « Acquisition automatique de lexiques sémantiques pour la recherche d'information », Thèse de doctorat, Université de Rennes 1, 2003.

[CLA 03b] CLAVEAU V., SÉBILLOT P., FABRE C., BOUILLON P., « Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus using Inductive Logic Programming », *Journal of Machine Learning Research, special issue on ILP*, vol. 4, 2003, p. 493-525.

[CLA 04a] CLAVEAU V., L'HOMME M.-C., « Discovering Specific Semantic Relationships between Nouns and Verbs in a Specialized French Corpus », *Proceedings of the 3rd Workshop on Computational Terminology, CompuTerm'04*, Genève, Suisse, 2004.

[CLA 04b] CLAVEAU V., SÉBILLOT P., « Extension de requêtes par lien sémantique nom-verbe acquis sur corpus », *Actes de la 11ème conférence de Traitement automatique des langues naturelles, TALN'04*, Fès, Maroc, 2004.

[CUS 00] CUSSENS J., DŽEROSKI S., Eds., *Learning Language in Logic*, LNAI, Springer Verlag, 2000.

[DAI 94] DAILLE B., « Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques », Thèse de doctorat, Université Paris 7, France, 1994.

[DRO 03] DROUIN P., « Term-extraction Using Non-technical Corpora as a Point of Leverage », *Terminology*, vol. 9, n° 1, 2003, p. 99-115.

[FAB 96] FABRE C., « Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations », Thèse de doctorat, Université de Rennes 1, France, 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, Symposium on Advances in Intelligent Data Analysis AIDA*, Rochester, NY, États-Unis, 1999.

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

[GAR 00] GARCIA D., AUSSENAC-GILLES N., COURCELLE A., « Exploitation, pour la modélisation, des connaissances causales repérées par COATIS dans les textes », CHARLET J., ZACKLAD M., KASSEL G., BOURIGAULT D., Eds., *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, Eyrolles, 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 95] GREFENSTETTE G., TEUFEL S., « Corpus-Based Method for Automatic Identification of Support Verbs for Nominalizations », *7th Conference of European Chapter of the Association for Computational Linguistics, EACL 95*, Dublin, Irlande, 1995.

[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*, Montréal, Canada, 1997.

[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 », *Boston Studies in the Philosophy of Science*, vol. 104, 1989, Kluwer Academic Publisher.

[HEA 92] HEARST M. A., « Automatic Acquisition of Hyponyms from Large Text Corpora », *14th International Conference on Computational Linguistics, COLING'92*, Nantes, France, 1992.

[HEA 98] HEARST M. A., « Automated Discovery of WordNet Relations », FELLBAUM C., Ed., *WordNet : an Electronic Lexical Database*, chapitre 5, p. 131–151, MIT Press, Cambridge, MA, États-Unis, 1998.

[JAC 01] JACQUEMIN C., *Spotting and Discovering Terms through NLP*, MIT Press, Cambridge, MA, États-Unis, 2001.

[JON 99] JONES R., MCCALLUM A., NIGAM K., RILOFF E., « Bootstrapping for Text Learning Tasks », *IJCAI-99 Workshop on Text Mining : Foundations, Techniques and Applications*, Stockholm, Suède, 1999.

[JOU 95] JOUIS C., « Seek, un logiciel d'acquisition de connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe », *6es Journées Acquisition, Validation, JAVA'95*, Grenoble, France, 1995.

[JOU 97] JOUIS C., BISKRI I., DESCLES J.-P., PRIO F. L., MEUNIER J.-G., MUSTAPHA W., NAULT G., « Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques », *1es Journées scientifiques et techniques du réseau francophone de l'ingénierie de la langue de l'AUPELF-UREF*, Avignon, France, 1997.

[KOH 95] KOHAVI R., « A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection », *14th International Joint Conference on Artificial Intelligence, IJCAI 95*, Montréal, Canada, 1995.

[MAN 99] MANNING C. D., SCHÜTZE H., *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, MA, États-Unis, 1999.

[MOR 98] MORIN E., « PROMÉTHÉE : un outil d'aide à l'acquisition de relations sémantiques entre termes », *Traitement Automatique des Langues Naturelles, TALN'98*, Paris, France, 1998.

[MOR 99] MORIN E., « Extraction de liens sémantiques entre termes à partir de corpus de textes techniques », Thèse de doctorat, Université de Nantes, France, 1999.

[MUG 94a] MUGGLETON S., « Bayesian Inductive Logic Programming », *7th Annual ACM Conference on Computational Learning Theory*, New York, NY, États-Unis, 1994.

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

[PEA 02] PEARCE D., « A Comparative Evaluation of Collocation Extraction Techniques », *3rd International Conference on Language Resources and Evaluation, LREC 02*, Las Palmas de Gran Canaria, Espagne, 2002.

[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, France.

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

[PUS 95] PUSTEJOVSKY J., *The Generative Lexicon*, The MIT Press, Cambridge, MA, États-Unis, 1995.

[PUS 97] PUSTEJOVSKY J., BOGURAEV B., VERHAGEN M., BUITELAAR P., JOHNSTON M., « Semantic Indexing and Typed Hyperlinking », *American Association for Artifical Intelligence Conference, Spring Symposium, NLP for WWW.*, Stanford University, CA., États-Unis, 1997, p. 120–128.

[RIL 99] RILOFF E., JONES R., « Learning Dictionaries for Information Extraction by Multi-level Bootstrapping », *16th National Conference on Artificial Intelligence, AAAI*, Orlando, FL, États-Unis, 1999.

[SRI 01] SRINIVASAN A., « The ALEPH manual », 2001.

[WIL 96] WILKS Y., STEVENSON M., « The Grammar of Sense : is Word-Sense Tagging much more than Part-of-Speech Tagging ? », Rapport de recherche, 1996, University of Sheffield, Royaume-Uni.

[YAR 95] YAROWSKY D., « Unsupervised Word Sense Disambiguation Rivaling Supervised Methods », *33rd Annual Meeting of the Association for Computational Linguistics, ACL*, Cambridge, MA, États-Unis, 1995.