TALN 2002, Nancy, 24-27 juin 2002

Évaluer l'acquisition semi-automatique de classes sémantiques

Thierry Poibeau, Dominique Dutoit et Sophie Bizouard

(1) Thales et LIPN
Domaine de Corbeville, 91404 Orsay, France
thierry.poibeau@thalesgroup.com

(2) Memodata et CRISCO
17, rue Dumont d'Urville, 14 000 Caen, France
memodata@wanadoo.fr

(3) Crim/Inalco
2, rue de Lille, 75007 Paris, France
sophie.bizouard@club-internet.fr

Résumé – Abstract

Cet article vise à évaluer deux approches différentes pour la constitution de classes sémantiques. Une approche endogène (acquisition à partir d'un corpus) est contrastée avec une approche exogène (à travers un réseau sémantique riche). L'article présente une évaluation fine de ces deux techniques.

This paper evaluates two different approaches for the elaboration of semantic classes. An endogeneous approach (corpus-based learning) is contrasted with a exogeneous one (the use of a large semantic network). The two techniques are finally evaluated.

Keywords – Mots Clés

Classes sémantiques, Évaluation, Ressources, Réseau sémantique
Semantic classes, Evaluation, Resources, Semantic network

1 Introduction : ressources générales *vs.* ressources spécifiques

Cet article vise à évaluer deux approches pour la constitution de classes sémantiques. Nous nous plaçons dans la perspective d'une application d'extraction d'information,

pour laquelle la notion de classe sémantique est primordiale[1]. Dans cette introduction, nous reprenons les deux principales approches généralement considérées : l'accès direct à des ressources générales et l'élaboration dynamique de ressources à partir du corpus de travail.

Les ressources générales ont été fortement critiquées ces dernières années. Les principales objections sont rappelées dans la thèse de D. Dutoit (2000) :

- Les ressources linguistiques générales (en particulier les réseaux sémantiques larges et hors-domaine) sont difficilement utilisables en contexte car elles sont abstraites et encodent rarement les traits propres à un domaine visé, qui sont pourtant essentiels à l'analyse.
- Même si l'information pertinente a été modélisée, elle est noyée au milieu de nombreuses informations non pertinentes pour la tâche et elle est donc difficilement exploitable.
- L'information à modéliser est si importante que la tâche est infinie et donc vouée à l'échec (Victorri, 1998).

L'utilisabilité d'une ressource vient en grande partie de la façon dont elle est structurée. De ce point de vue, il faut noter que la plupart des critiques examinées ci-dessus s'adressent en premier lieu au réseau WORDNET de Princeton (Fellbaum, 1998), qui souffre certainement d'un manque de structuration pour être exploitable de façon optimale dans une visée applicative[2]. Sur un plan plus théorique, la compréhension est un phénomène qui met en jeu un réseau de connaissances *a priori* et l'usage d'un réseau sémantique, même incomplet, permet de se rapprocher de cet état de fait. Nul mécanisme de compréhension ne peut se contenter d'une analyse purement endogène, où toute la connaissance viendrait du contexte immédiat.

Il n'en demeure pas moins que certains chercheurs, refusant les arguments ci-dessus, se sont plutôt tournés vers l'acquisition de classes sémantiques à partir de corpus (Hirschmann *et al*, 1975) (Grishman et Sterling, 1994). Les méthodes d'acquisition à partir de corpus reposent pour la plupart sur une « hypothèse distributionnaliste » (Harris, 1951) : à l'intérieur d'un sous-langage, le repérage de structures syntaxiques régulières permet de mettre en évidence des familles de mots apparaissant dans des contextes communs. Ces familles de mots servent à former des classes sémantiques.

L'approche est séduisante, mais les résultats sont nuancés. La question de la taille du corpus, de sa nature et de sa variabilité se pose en effet de façon cruciale. Cet aspect est relativement peu abordé dans les comptes rendus d'expériences, qui portent souvent sur

---

[1] Nous détaillerons ici l'évaluation des classes sémantiques en elles-mêmes, en prenant en compte leur degré de recouvrement notamment. Par manque de place, nous ne pouvons ici présenter une évaluation et une discussion de leur influence sur le processus d'extraction lui-même. Sur ce sujet, voir [Poibeau, 2002].

[2] Depuis, un modèle WORDNET plus complexe a été mis au point pour améliorer certains des points faibles du modèle initial : il s'agit de EUROWORDNET, qui a été conçu lors de la création de réseaux sémantiques multilingues pour plusieurs langues européennes (Vossen, 1998).

des corpus réguliers ou techniques (par exemple, Nazarenko *et al.* (1997) proposent une expérience sur le domaine médical, Faure (2000) sur des recettes de cuisine). Plus fondamentalement, la plupart des études mentionnent que les classes obtenues automatiquement ne sont pas complètement satisfaisantes (Nazarenko *et al.*, 2001) : elles ne peuvent pas être injectées directement dans une application. Plusieurs raisons bien connues expliquent cet état de fait :

- Même à l'intérieur d'un sous-langage, un schéma syntaxique peut être ambigu.
- Les mots apparaissant dans un contexte donné entretiennent entre eux des relations plus ou moins étroites[3].
- Certains mots sont très généraux et permettent l'induction de classes sans grande pertinence.

Il est donc nécessaire d'injecter des connaissances dans le système d'acquisition pour en améliorer les résultats et le rendre utilisable. L'apprentissage devient alors supervisé, comme dans l'expérience de Nazarenko *et al.* (2001) qui proposent d'adapter la nomenclature médicale SNOMED au corpus MENELAS. Les auteurs projettent les catégories de SNOMED sur le résultat d'une analyse distributionnelle, ce qui contribue à affiner les classes obtenues automatiquement. Morin et Jacquemin (1999) procèdent de manière similaire pour spécialiser un thésaurus (AGROVOC) en fonction d'un corpus.

Le débat est donc moins tranché qu'il pourrait le paraître de prime abord entre ressources générales et ressources spécialisées. Nous proposons donc de reprendre cette question ici d'un point de vue applicatif, en comparant deux approches pour l'acquisition semi-automatique de classes sémantiques. Nous évaluons d'abord le système ASIUM (Faure, 2000) proposant une acquisition automatique à partir de corpus puis une phase de validation interactive (section 3). Nous effectuons ensuite la même expérience avec le réseau sémantique de MEMODATA (Dutoit, 2000), qui constitue à notre connaissance le réseau sémantique le plus riche à l'heure actuelle pour le français (section 4). Nous présentons ensuite des conclusions sur le processus d'évaluation en lui-même (section 5).

2 Cadre de l'expérience

Les expériences décrites ont été menées sur un corpus de dépêches financières. 300 dépêches traitant de rachat et de vente d'entreprises ont été retenues (corpus FirstInvest[4]). Cet ensemble de dépêches a été divisé en deux parties : 57 dépêches pour le corpus de test, 243 autres pour le corpus d'entraînement. Un système d'extraction sur ce domaine a été développé. Les ressources élaborées manuellement incluaient des

[3] Dans un corpus de cuisine, le verbe *conserver (au)* peut être suivi de *frais, froid, réfrigérateur, sec,* etc [Faure, 2000]. Même pour l'expert, le découpage n'est pas évident : faut-il considérer que, dans le corpus, *froid* est synonyme de *réfrigérateur* ? Les regroupements dépendent alors du degré de finesse recherché et peuvent difficilement être déterminés automatiquement.

[4] Firstinvest.com est un site financier sur Internet. La fourniture de ce corpus par FirstInvest a rendu ce travail possible.

classes sémantiques qui ont constitué la référence pour les expériences réalisées ensuite.

3 Acquisition automatique de familles sémantiques par apprentissage symbolique interactif

Nous avons entrepris d'évaluer l'apport d'un outil d'apprentissage de connaissances à partir de données textuelles pour élaborer les ressources nécessaires à un système d'extraction. Nous donnons ci-dessous une rapide description du système que nous avons utilisé, ASIUM (pour une description plus complète, cf. (Faure et Nédellec, 1998), (Faure, 2000)).

3.1 Le système d'apprentissage ASIUM

Le système ASIUM procède à un apprentissage symbolique non supervisé à partir de textes annotés syntaxiquement. ASIUM a pour objectif d'aider l'utilisateur à créer un ensemble de classes sémantiques pertinentes par rapport à la tâche et à les formaliser en ontologies. Ces ontologies doivent permettre de faire apparaître clairement les relations existantes entre classes. L'originalité de la méthode réside essentiellement dans l'étape de validation interactive. L'approche repose sur une analyse distributionnelle, permettant de générer des classes de mots apparaissant dans des contextes similaires (les *classes de base*). La mesure de similarité d'ASIUM permet de calculer le recouvrement entre classes et de proposer le regroupement de classes ayant une similarité supérieure à un seuil fixé par l'expert. Les classes de base sont successivement agrégées par une méthode coopérative, ascendante en largeur d'abord afin de former les concepts de l'ontologie niveau par niveau. Ce processus de catégorisation ne se contente pas d'identifier des listes de noms apparaissant dans des contextes similaires, mais il augmente ces listes par *induction*. En effet, le rassemblement de deux classes de base (`C1` et `C2`) trouvées après deux contextes différents `CTXT1` et `CTXT2` permet d'autoriser les séquences non attestées `CTXT1 C2` et `CTXT2 C1`. Par exemple, imaginons que les mots *acquisition* et *fusion* aient été regroupés au sein d'une même classe. Si *procéder à une acquisition* est attesté, le système va induire que *procéder à une fusion* est une structure valide, même si elle n'est pas attestée en corpus. Le résultat est une généralisation des connaissances trouvées dans le corpus. ASIUM requiert donc, à ce stade, l'intervention d'un expert du domaine pour vérifier les classes une à une, classées par mesure de similarité décroissante, jusqu'à atteindre le seuil en-deçà duquel les classes ne sont plus présentées à l'expert.

3.2 Apprentissage supervisé ou non supervisé ?

Les classes de base fournies par ASIUM nécessitent un important travail de validation. Du fait de la faible taille du corpus disponible, le seuil retenu pour les classes de base est obligatoirement bas et les rapprochements entre classes s'opèrent sur un nombre d'éléments communs relativement peu important.

Deux expériences ont été réalisées. Dans la première expérience, ASIUM est utilisé tel quel. Les rapprochements entre classes de base s'effectuent en mode non supervisé[5] : ils sont uniquement fondés sur le score obtenu entre les deux classes. Cette stratégie oblige à valider ou à refuser un nombre important de classes qui ont des éléments communs mais qui ne sont pas pertinentes pour le domaine. Une deuxième expérience a alors utilisé une fonctionnalité d'ASIUM permettant de focaliser la recherche sur les seules classes comprenant des éléments du domaine, définis d'après un ensemble de mots placés dans un fichier jouant le rôle de filtre. Par exemple, pour la recherche de la classe `Opération d'achat`, le filtre comprenait les mots *achat*, *fusion, scission* ; pour la classe `Entreprise`, les mots *entreprise*, *société*, *filiale*. Cette stratégie débouche sur un apprentissage de type supervisé[6], ce qui améliore notablement la pertinence du résultat, à condition de définir à chaque fois un ensemble de mots pertinents. Suivant la classe sémantique que l'analyste cherche à modéliser, cet ensemble de mots n'est pas le même. Il y a donc là un premier facteur de subjectivité dans l'évaluation : la définition du filtre est une opération subjective qui influe notablement sur les résultats.

3.3 Critères pour l'élaboration des classes

La modélisation des classes nécessite un travail manuel significatif. Même avec un fichier jouant le rôle de filtre, le nombre de classes de base proposées est important, puisqu'il a varié de 41 à 576 suivant le nombre d'éléments contenus dans le filtre. La plupart du temps, la classe de base ASIUM la plus proche de la classe sémantique visée contient moins de 30 % des éléments désirés ; parallèlement, il est rare que plus de 30 % d'une classe de base soit jugé pertinent. Le travail de nettoyage des classes et le nombre d'agrégations de classes nécessaires s'en trouvent d'autant accrus. Il n'est pas rare de trouver des classes sémantiques réparties sur 20 classes de base, nécessitant autant d'agrégations. Ce point peut poser problème lors de la mise en œuvre du système : il faut que l'expert accepte de passer du temps à valider des classes initialement peu représentatives.

3.4 Mesurer l'apport de l'apprentissage pour l'acquisition de ressources

Le but de l'expérience est de tenter de repérer semi-automatiquement, grâce à ASIUM, les classes sémantiques qui sont utiles pour l'extraction. Dans le cadre de l'expérience sur le corpus FirstInvest, l'évaluation a porté sur une comparaison des items figurant dans la classe préalablement définie par l'expert avec ceux obtenus par apprentissage.

---

[5] A ce stade, l'utilisateur n'est pas dans la boucle : il intervient juste pour valider les propositions faites par le système.

[6] Cette stratégie fait suite à l'utilisation d'Asium dans des expériences d'acquisition de ressources pour l'extraction. De façon analogue, Riloff et Jones (1999) proposent une approche fondée sur un ensemble de mots-clés initiaux dits *seed words* pour amorcer l'apprentissage de classes sémantiques. Ces expériences ont montré qu'il était nécessaire de mettre au point un moyen de superviser l'apprentissage, contrairement à ce qui est recherché dans l'acquisition d'ontologies, qui porte généralement sur tout le texte. L'extraction ne s'intéresse, rappelons-le, qu'à une partie minime du texte.

L'analyse d'ASIUM met en jeu la fréquence d'apparition des éléments : de fait, presque tous les éléments ayant une influence majeure sur le résultat ont été trouvés. Les « oublis » portent sur des éléments rares, influençant peu le résultat final, comme *échange*, *émission*, *transaction*. Il faut toutefois noter que certains éléments comme *reprise* ou *vente* n'ont pas été trouvés, bien qu'il s'agisse de mots fréquents[7]. L'analyse distributionnelle permet en outre de retrouver quelques éléments pertinents pour la tâche, qui n'avaient pas été retenus lors du travail d'évaluation manuel des ressources (*désengagement*, *regroupement*, *scission*). Bien évidemment aucun élément n'apparaissant pas dans le corpus d'entraînement ne peut figurer dans une classe proposée par ASIUM.

Une fois affinées et validées, les classes définies semi-automatiquement permettent d'obtenir des résultats globaux relativement bons. Nous avons affaire à un corpus homogène : malgré sa taille réduite (243 textes, soit 45 334 mots), le corpus d'entraînement permet d'obtenir des performances correctes sur le corpus de test (80 % de précision et 60 % de rappel en moyenne suivant les classes sémantiques visées). L'utilisation d'ASIUM, accompagnée d'un important travail de validation, permet de bien couvrir le corpus d'entraînement.

L'outil permet d'établir conjointement des listes de noms et de verbes associés. Il offre donc un gain de temps par rapport à la lecture du corpus. L'évaluateur estime avoir passé 10 heures à définir des classes sémantiques avec ASIUM, à comparer avec les 50 heures consacrées à lire le corpus lors du développement manuel des ressources. Le filtrage d'après un ensemble de mots clés permet d'obtenir des classes pertinentes dès le début de la validation, ce qui offre un avantage important par rapport à l'expérience de Faure et Poibeau (2000) qui reposait sur un apprentissage non supervisé.

4 Utilisation d'une ressource linguistique générale : le réseau sémantique de Memodata

Afin de contraster l'expérience faite avec ASIUM, nous avons décidé de procéder à la même manipulation en partant d'un réseau sémantique de la langue générale. Nous avons retenu le réseau de MEMODATA, qui a développé depuis plus de 10 ans un outil très complet, le DICTIONNAIRE INTEGRAL, accompagné de fonctionnalités accessibles via une interface de programmation (API, *Application Programming Interface*) (Dutoit, 2000).

4.1 Le réseau sémantique : le DICTIONNAIRE INTEGRAL et les outils associés

Le DICTIONNAIRE INTEGRAL est le nom du réseau sémantique développé par MEMODATA. Ce réseau est fondé sur la notion de mots-sens, c'est-à-dire sur les

[7] Bien qu'ils soient fréquents, ces mots ne sont pas retenus par l'analyse car ils se trouvent dans des contextes différents des mots supportant la notion d'achat.

différents sens associés à un mot (une chaîne de caractère). Il est riche d'environ 186 000 mots-sens : il est donc comparable par la taille à WORDNET (Fellbaum, 1998), mais il en diffère notablement par la façon dont l'information y est structurée. En effet, la décomposition n'est pas seulement fondée sur une relation « est-un » et sur des présupposées psychologiques, mais aussi sur une décomposition des mots-sens en sèmes (analyse componentielle) structurés par des liens typés.

Le réseau repose sur de nombreuses relations entre des éléments de catégories syntaxiques différentes, à l'instar de EUROWORDNET (Vossen, 1998). Cet aspect est bien évidemment primordial pour la tâche qui nous intéresse, puisqu'il nous faut par exemple trouver aussi bien *achat* que *acheter* pour refléter la notion d'`acquisition`. La couverture du DICTIONNAIRE INTEGRAL est plus large que celle du réseau sémantique au format EUROWORDNET existant pour le français, une partie de l'information contenue dans le EUROWORDNET provenant d'ailleurs des bases de MEMODATA (le reste du réseau EUROWORDNET français ayant été produit par l'Université d'Avignon, responsable du projet pour la France).

4.2 Critères pour l'élaboration des classes

Pour élaborer des classes sémantiques à partir d'un réseau comme celui de Memodata, l'utilisateur doit choisir un mot-clé du domaine visé comme point d'entrée. La classe émerge alors progressivement en suivant des liens dans ce réseau.

L'expérience a pris en compte tous les mots reliés à un mot-clé par un lien de type *générique*, *spécifique* ou *synonyme*. La classe lexicale ainsi obtenue est affinée de manière à obtenir une classe sémantique relativement homogène et pertinente (il s'agit essentiellement d'enlever les mots non pertinents pour la tâche ou le domaine). Ainsi, en partant de *achat*, on accède à une classe comprenant les mots suivants : *abonnement, accaparement, acheter, acquérir, acquisition, appropriation, commande, prise de contrôle, prise de participation, préemption, téléachat*… Dans le contexte de rachat d'entreprise, il est douteux que *abonnement* ou *téléachat* soient pertinents. Ces mots doivent être éliminés à la main.

A la différence de l'expérience avec ASIUM, l'utilisateur n'a pas d'information sur la représentativité d'un mot par rapport à son corpus : il ignore sa fréquence et ses contextes d'apparition ; le retour « manuel » au corpus est lourd et fastidieux. Nous verrons dans la section suivante que la représentativité des résultats s'en ressent. L'arrêt du processus d'acquisition intervient quand les liens examinés ne rapportent quasiment plus de mots pertinents. Les mots les plus éloignés retenus sont généralement séparés de seulement 2 à 3 nœuds du mot-clé initial. Le réseau est homogène : les spécifiques des synonymes du mot-clé initial couvrent souvent quasi parfaitement les spécifiques du mot-clé initial[8]. Dans les faits, les relations suivies

[8] Ceci est moins évident qu'il n'y paraît : les liens de synonymie dans les réseaux sont rarement réciproques (un lien de synonymie peut exister de A vers B, mais pas de B vers A), contrairement à une idée reçue.

ramènent rapidement soit des ensemble de mots déjà trouvés, soit des ensembles peu pertinents. Le processus d'acquisition s'arrête alors.

4.3 Mesurer l'apport de ressources générales pour l'acquisition de classes sémantiques

L'évaluation de l'apport du DICTIONNAIRE INTEGRAL pour la définition des classes sémantiques du domaine se rapproche partiellement de celle qui a été mise en œuvre pour évaluer ASIUM. Il s'agit en effet de comparer la couverture de classes obtenues à partir de l'outil de MEMODATA avec la couverture des classes définies manuellement. On constate toutefois quelques différences : l'outil de MEMODATA ne nécessite pas de travail de modélisation ni de corpus d'entraînement. Plusieurs facteurs de subjectivité apparaissent aussi de manière flagrante dans l'évaluation. Ainsi, l'interrogation de l'outil de MEMODATA se fait à partir d'un mot ou d'un ensemble de mots-clés. Le point de départ (mot-clé servant de support à l'interrogation) a une influence majeure sur les résultats obtenus, du fait qu'un mot peut être plus ou moins bien relié aux mots de sens proche, que le réseau lexical est plus dense dans certains domaines, etc.

On observe qu'en interrogeant le DICTIONNAIRE INTEGRAL avec 3 à 5 mots seulement, il est possible de reconstituer intégralement n'importe laquelle des classes sémantiques sur lesquelles a porté l'évaluation. Toute la difficulté consiste évidemment à bien choisir les mots-clés initiaux (comme *achat* dans l'exemple du paragraphe précédent), qui sont déterminants pour la qualité du résultat final. On retrouve ici le même facteur de subjectivité que celui qui préside à l'élaboration des filtres d'ASIUM. Le petit nombre d'éléments initiaux du DICTIONNAIRE INTEGRAL à activer pour reconstituer les classes sémantiques visées montrent toutefois l'homogénéité des classes sémantiques définies par MEMODATA. Pour éviter la subjectivité qu'entraîne le choix d'une combinaison de mots-clés, nous avons voulu faire l'évaluation en ne prenant en compte qu'un seul mot comme point de départ, à savoir le mot « *achat* ». La classe obtenue en activant simplement les liens directs de « *achat* » puis en validant les résultats permet de découvrir plus de 75 % de la classe sémantique visée, si nous ne tenons pas compte de l'effectif dans le corpus de test des mots composant la classe. 60 % des éléments proposés par le DICTIONNAIRE INTEGRAL ont été retenus. Une large partie de la classe ainsi obtenue ne figurait pas dans la classe élaborée manuellement (il y a en moyenne 40 % d'éléments nouveaux dans les classes obtenues à partir du DICTIONNAIRE INTEGRAL, comparées aux classes définies manuellement). Autrement dit, le réseau de MEMODATA fourni beaucoup d'éléments auxquels un analyste n'aurait pas pensé spontanément.

Ces résultats sont toutefois trompeurs et sont très différents si nous raisonnons en termes d'effectif. Des éléments clés de la classe sémantique visée n'ont pas été trouvés. Les chiffres tenant compte de l'effectif révèlent que la classe obtenue avec le DICTIONNAIRE INTEGRAL ne couvre que 45 % de la classe visée, et les nouveaux éléments proposés ne contribuent qu'à hauteur de 7 % à l'amélioration de la couverture. Autrement dit, de nombreux termes proposés *pourraient* être pertinents, mais ceux-ci ne sont *pas* présents dans le corpus de test. Nous retiendrons que, globalement, les classes proposées par le DICTIONNAIRE INTEGRAL sont de meilleure qualité que celles proposées par ASIUM et elles couvrent un spectre plus large

d'éléments, du fait qu'elles ne sont pas limitées au corpus d'entraînement. La médaille a hélas un revers : des éléments clés du corpus sont oubliés (30 % d'éléments non couverts représentent en fait 55 % des occurrences en corpus). Certains de ces éléments sont des éléments fréquents, ce qui contribue à faire largement baisser le rappel global.

5 Les outils d'acquisition de ressources, une évaluation difficile

Les outils d'extraction offrent des possibilités d'évaluation relativement claires : une référence peut généralement être établie à la main, et les résultats de l'extraction sont comparés à cette référence. Il en va tout autrement quand il s'agit d'acquisition de ressources. Les ressources définies manuellement constituent une référence relative, que l'on sait largement imparfaite. L'évaluation ne repose donc pas sur une réalité objective et incontestable. Les difficultés deviennent encore plus grandes quand on sait que l'élaboration de ressources ne peut constituer une tâche isolée et abstraite. En effet, le travail de modélisation linguistique exige une connaissance du domaine et du corpus traité. Cette connaissance permet à l'utilisateur de guider le système d'apprentissage en définissant les « bons » filtres ou les « bons » mots-clés. Mais, comme toute connaissance humaine, elle est difficile à modéliser, comporte une part de non-dit et d'implicite.

Cette part de subjectivité est inhérente à la tâche et ne doit pas être masquée lors de l'évaluation. On obtient alors des résultats moins lisibles et moins clairs que lors d'une évaluation de type « boîte noire ». Ce manque de netteté est souvent compensé par une intuition de l'expert qui développe ses propres stratégies d'analyse et d'utilisation des outils à sa disposition. Il est donc nécessaire de proposer des guides lors de la mise en place de tels outils : guides des fonctionnalités, mais aussi et surtout guides méthodologiques explicitant les stratégies d'analyse possibles en fonction de la taille du corpus, de son homogénéité, de sa représentativité et du domaine considéré.

6 Conclusion

Dans cet article, nous avons procédé à l'évaluation de deux sources de connaissances pour l'acquisition de classes sémantiques : le corpus lui-même et un réseau sémantique général. Les résultats obtenus sont contrastés et nous avons montré que les deux sources de connaissances ne permettent pas d'obtenir les même résultats. L'acquisition à partir du corpus est plus précise, tandis que le recours à un réseau sémantique permet de couvrir plus « large ». La complémentarité de ces deux sources de données reste maintenant à être mieux exploitée au sein d'une système coopératif.

7 Bibliographie

DUTOIT D. – *Quelques opérations texte → sens et sens → texte utilisant une sémantique linguistique universaliste apriorique*. Thèse de Doctorat en Informatique, Université de Caen, 2000.

FAURE D. – *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de texte : le système ASIUM.* Thèse de Doctorat en Informatique, Université Paris 11–Orsay, 2000.

FAURE D., NÉDELLEC C. – « ASIUM: Learning subcategorization frames and restrictions of selection ». In *Proceedings of the ECML'98 Workshop on Text Mining*, Chemnitz, 1998.

FAURE D. et POIBEAU T. – « Extraction d'information utilisant INTEX et des connaissances sémantiques apprises par ASIUM, premières expérimentations ». In *Actes du 12ème Congrès de Reconnaissance des Formes et Intelligence Artificielle* (RFIA'2000), Paris, 2000, pp. 91–100.

FELLBAUM S. (éd.)– *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, 1998.

GRISHMAN R. et STERLING J. – « Generalizing automatically generated selectional patterns ». In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING'94), Tokyo, 1994, pp. 742–747.

HARRIS Z. – *Structural Linguistics.* University of Chicago Press, Chicago, 1951.

HIRSCHMAN L., GRISHMAN R. et SAGER N. – « Grammatically-based automatic word class formation ». *Information Processing and Management*, vol. 11, 1975, pp. 39–57.

MORIN E.et JACQUEMIN C. – « Projecting corpus-based semantic links on a thesaurus ». In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL'99), Université du Maryland, 1999, pp. 389–396.

NAZARENKO A., ZWEIGENBAUM P., BOUAUD J. et HABERT B. – « Corpus-based identification and refinement of semantic classes ». *Journal of the American Medical Informatics Association*, vol. 4, 1997, pp. 585–589.

NAZARENKO A., ZWEIGENBAUM P., HABERT B. et BOUAUD J. – « Corpus-based extension of a terminological semantic lexicon ». In BOURIGAULT D., JACQUEMIN C. et L'HOMME M.-C. (éds.) – *Recent advances in Computational Terminology.* John Benjamins, Amsterdam, 2001, pp. 327–352.

POIBEAU T. – *Extraction d'information à base de connaissances hybrides*. Thèse de Doctorat en Informatique, Université de Villetaneuse, 2002.

RILOFF E. et JONES R. – « Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping ». In *Proceedings of the 16th National Conference on Artificial Intelligence* (AAAI'99), Orlando, 1999, pp. 474–479.

VICTORRI B. – « La construction dynamique du sens ». In *Actes du 11ème Congrès de Reconnaissance des Formes et Intelligence Artificielle* (RFIA'1998), Clermont Ferrand, 1998, pp. 15–29.

VOSSEN, P. (éd.) – *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.