Adaptation d'une Ressource Prédicative pour l'Extraction d'Information

Aurélien Bossard, Thierry Poibeau

Adaptation d'une ressource prédicative pour l'extraction d'information

Aurélien Bossard et Thierry Poibeau [1]
Laboratoire d'Informatique de Paris-Nord

Abstract

In this article, we present a method aiming at building a resource for an information extraction task, from an already existing French predicative lexical resource. We point out the advantages and drawbacks of two predicative resources we worked with: the LADL tables and Volem. We present the reasons why we finally selected Volem as the most interesting resource for the task. Thereafter, we make a list of the needs an information extraction task implies, and how we include missing information in the resource we selected. We evaluate the resource completed by those missing informations, using it in an information extraction task.

Keywords : Predicative schemas, information extraction, lexico-syntactic patterns.

1. Introduction

Cet article vise à montrer la façon dont nous avons procédé pour contribuer à une ressource lexicale pour le français, utile à des fins d'extraction d'information, en nous servant de ressources déjà existantes. La création d'une ressource pour le français assez complète pour couvrir de larges champs d'applications en traitement des langues (TAL) apparaît aujourd'hui comme un enjeu majeur. En effet, en raison du manque de ressources, deux approches sont possibles afin de réaliser des applications du type « extraction d'information » : l'apprentissage automatique et/ou la création de règles à la main. Aucune de ces deux méthodes n'est parfaite : l'approche manuelle risque d'entraîner des ressources incomplètes, inconsistantes et pose le problème de la réutilisabilité. Les approches automatiques présentent des risques d'erreur qui doivent être corrigés manuellement et reposent le plus souvent sur des corpus annotés, qui ne sont pas toujours disponibles.

L'approche que nous défendons ici est fondée sur la notion de schéma prédicatif. Les éléments pertinents pour l'extraction sont ceux qui se situent autour d'une relation sémantique, portée par un nom prédicatif ou, plus souvent, par un verbe. L'étude des schémas prédicatifs permet d'attribuer un rôle à chacun des arguments du prédicat ; l'étude de la syntaxe de la phrase permet de reconnaître ces arguments malgré les variations de surface. Nous testons notre approche dans le cadre d'une application d'extraction d'information portant sur des rachats d'entreprise. La tâche consiste par exemple à donner une représentation (sémantique) identique pour les trois phrases suivantes :

[1] Laboratoire d'Informatique de Paris-Nord
CNRS UMR 7030 et Université Paris 13
99, avenue J.-B. Clément, F-93430 Villetaneuse
firstname.lastname@lipn.univ-paris13.fr

- CPI rachète Fulmar.
- CPI a racheté Fulmar à son PDG pour 50 millions d'euros.
- CPI a indiqué avoir racheté Fulmar.

Dans chacune de ces trois phrases, nous pouvons identifier un acheteur : *CPI* et une entreprise achetée : *Fulmar*, éléments auxquels peuvent s'ajouter des données sur le montant de la transaction, le vendeur, la date de la transaction... Ce type d'applications a déjà été développé, y compris pour le français (Poibeau, 2003). Notre but est ici de l'envisager avec un nouveau regard, en nous focalisant sur des ressources génériques et sémantiquement riches. Plutôt que de développer des ressources de manière *ad hoc*, nous cherchons à caractériser l'intérêt des données déjà existantes pour le français.

Les questions que nous nous sommes posées sont les suivantes : quelles sont les informations qu'une ressource lexicale doit encoder pour que l'on puisse arriver à un tel résultat ? Est-il possible, avec les ressources existantes, de créer une telle ressource ? Quel réel intérêt aurait une telle ressource (précision de l'extraction, rappel, automatisation...) ? Dans un premier temps, nous présentons quelques ressources lexicales existantes pour le français. Dans une seconde partie, nous caractérisons plus en détail notre sujet d'étude avant d'aborder, dans une troisième partie, les expériences réalisées. Nous présentons ensuite les résultats et, dans une dernière partie, les conclusions de notre recherche.

2. Choix d'une ressource

L'anglais dispose aujourd'hui de trois ressources à large couverture encodant d'une manière ou d'une autre la notion de schéma prédicatif : VerbNet (Kipper-Schuler, 2003), PropBank (Palmer *et al.*, 2005) et FrameNet (Fillmore *et al.*, 2003). Ces trois ressources sont fondées sur des approches différentes: approche syntaxique pour VerbNet et PropBank, et approche sémantique pour FrameNet (Pitel, 2006). De nombreuses applications ont été développées autour de FrameNet, comme la désambiguïsation sémantique (Fillmore & Baker, 2001), (Lowe *et al.*, 1997), mais aussi l'extraction d'information. Des recherches ont été menées pour utiliser conjointement ces trois ressources afin d'améliorer l'étiquetage sémantique (Giuglea & Moschitti, 2004).

Il existe beaucoup moins de richesse pour le français. Le Dictionnaire Explicatif et Combinatoire (DEC, `www.olst.umontreal.ca/decfr.html`) d'I. Melc'uk a été exclu car il n'offre pas une couverture suffisante pour la tâche. DicoValence (`bach.arts.kuleuven.be/dicovalence`) n'était quant à lui pas disponible au moment de l'étude mais mériterait sinon d'être pris en considération. Nous nous sommes alors focalisés sur deux ressources pour le français : Volem et les tables du LADL. Il s'agit dans cette partie d'expliquer le choix que nous avons fait concernant la ressource que nous avons utilisée.

2.1. Les Tables du LADL

Les Tables du LADL, aussi connues sous le nom de lexique-grammaire, ont été établies sous la direction de Maurice Gross. Elles regroupent 6000 verbes répartis dans des tables construites d'après des similitudes de comportement syntaxique. Chaque table du Lexique-Grammaire contient un certain nombre de propriétés, qui sont validées ou invalidées pour chacun des verbes qui y figure (matrice de + et de -). Les propriétés encodent des informations sur (Gross, 1975) :

- Les réalisations possibles des arguments (restrictions de sélection : arguments à trait "humain" ou "non-humain", argument de type abstrait...) ;
- Les propriétés syntaxiques du verbe ou de ses arguments (pronominalisation possibles des arguments, type de préposition pour chaque argument...) ;
- Les sous-catégorisations alternatives ;

- Les possibilités de redistributions (passif long, passif court...).

Les informations contenues dans les tables du LADL sont riches sur le plan syntaxique mais ne fournissent pas directement l'information dont nous avons besoin au plan sémantique. Les arguments sont typés par des restrictions de sélection (`humain`, `non-humain`, `objet concret`, `abstrait`...) mais ne mentionnent pas de rôle sémantique en tant que tel. L'information sur chaque argument est en outre codée sous forme de traits répartis sur plusieurs colonnes, ce qui rend les traitements difficiles : il est nécessaire d'effectuer un travail important de transformation pour rendre ces tables exploitables directement par des applications de TAL (Gardent *et al.*, 2005). Enfin, seule une partie des tables est actuellement disponible.

Il n'en reste pas moins que l'approche par automate patron et ensembles de contraintes (encodées au sein d'une matrice de traits) rend les données linguistiques lisibles, aisément modifiables et exploitables, ce qui est évidemment très intéressant pour notre approche. Les tables du LADL possèdent en outre une richesse et une couverture qui ne peut être ignorée.

2.2. Volem

Volem (Saint-Dizier *et al.*, 2002) est une ressource multilingue (français-espagnol-catalan). Les entrées sont des verbes : la ressource décrit leur comportement syntaxique et sémantique à travers la description des arguments et des schémas de sous-catégorisation. Cette ressource décrit à l'heure actuelle 1700 verbes.

Description du verbe : acheter

GRILLE THEMATIQUE :
[[inic(agent),dest],[th],[src]]
LCS :
ALTERNANCES :
caus_2np_pp , anti_pr_np , anti_pr_np_pp , pas_etre_part_np_2pp , pas_etre_part_np_pp , caus_2np , caus_refl_pr_2np , caus_np_pp , caus_support_np
WN :
[13,2,3], [13,3,1] , [13,3,8]
EXEMPLE :
Il a acheté ce livre à un brocanteur

Figure 1. L'entrée lexicale du verbe "acheter" dans Volem

Cette ressource est fondée sur une liste de rôles thématiques génériques mais néanmoins assez précis. Les différents rôles thématiques peuvent être combinés afin de décrire aux mieux les arguments d'un verbe (cf. figure 1).

Les principaux inconvénients de cette ressource sont:

- L'absence de gestion de la polysémie (les concepteurs de la ressource ont fait le choix de ne coder qu'un sens par verbe, correspondant à l'emploi le plus fréquent);
- La faible couverture de la ressource (1700 verbes);
- L'absence de description précise des schémas syntaxiques que représentent les différentes alternances utilisées dans Volem.

Volem a une couverture moindre que celle des tables du LADL. L'absence de gestion des rôles thématiques dans les tables du LADL constitue cependant un inconvénient de taille pour une tâche d'extraction d'information, qui demande des informations fines sur la nature des arguments. Nous avons donc choisi d'utiliser une combinaison de ces deux ressources : les données de Volem bien qu'incomplètes, correspondent relativement bien à nos besoins. Le format des tables du LADL est quant à lui mieux adapté à un traitement informatique.

Nous avons entrepris, dans un premier temps, de voir s'il était possible de de coder les informations contenues dans Volem sous forme de tables de type lexique-grammaire. Au-delà de cette étude, la fusion des données du LADL avec celle de Volem serait à envisager mais nous examinons ici une autre question, celle de l'enrichissement semi-automatique de Volem et la compilation de cette ressource sous forme de table du lexique-grammaire.

3. Méthode d'enrichissement de la ressource

Les informations contenues dans Volem ne sont pas suffisantes pour réaliser sans ajout une tâche d'extraction d'information. Le format XML de Volem n'est pas non plus directement exploitable. Nous avons donc modifié la ressource sur les points suivants, afin de pouvoir l'utiliser dans le cadre d'une tâche d'extraction d'information:

1. Codage de la ressource sous la forme d'une table de contraintes ;
2. Ajout des informations manquantes ;
3. Codage d'automates patrons.

3.1. Codage de la ressource sous la forme d'une table de contraintes

Nous voulons, à partir de Volem, créer des automates d'extraction au format Unitex : `www-igm.univ-mlv.fr/~unitex/`. Pour cela, nous avons besoin d'une ressource codée sous forme d'une table de contraintes. Nous avons donc créé un convertisseur permettant de passer automatiquement du format de Volem au format défini par le LADL. Nous utilisons une colonne par alternance de Volem, et validons l'alternance pour un verbe en mettant un "+" dans la case correspondante, et un "-" sinon. Pour encoder les informations sur les rôles thématiques, nous créons une colonne par argument du verbe et nous remplissons chacune d'elle avec la description du rôle thématique de l'argument, telle qu'elle est enregistrée au sein de Volem . Ainsi, la ressource est directement exploitable par des « graphes patrons » Unitex, qui exploitent les tables de contraintes.

3.2. Ajout des informations manquantes

La ressource en l'état n'encode toujours pas assez d'informations pour réaliser une extraction d'information précise. En effet, il lui manque encore:

1. le type d'auxiliaires requis ;
2. les différentes prépositions introduisant éventuellement un argument ;
3. les nominalisations (e.g. *rachat* pour *racheter*) ;
4. les adjonctions essentielles.

3.2.1. Les auxiliaires

Deux possibilités se sont offertes à nous pour ajouter les auxiliaires à la table de données que nous avions construite. Soit récupérer les auxiliaires depuis un dictionnaire, soit identifier l'auxiliaire d'un verbe grâce aux occurrences de celui-ci en corpus. Nous avons opté pour la deuxième solution, en utilisant un corpus constitué de textes journalistiques (dépêches financières et articles du journal *Le Monde*). L'acquisition est effectuée grâce à un automate qui reconnaît les différents verbes ainsi que les auxiliaires qui les accompagnent. Si l'auxiliaire "avoir" apparaît au moins une fois dans le texte pour un verbe donné, un "+" est ajouté à l'intersection de la ligne correspondant à ce verbe et de la colonne correspondant à l'auxiliaire "avoir".

3.2.2. L'ajout des prépositions

Nous avons déjà mentionné qu'une seule préposition est codée par argument au sein de Volem, alors que plusieurs prépositions peuvent apparaître pour certains verbes (*acheter à* ou *auprès de*). Nous avons donc réalisé un outil permettant d'ajouter à la table de données les différentes prépositions qui introduisent les arguments d'un verbe. Cet outil nécessite une validation des résultats par l'utilisateur.

Le système est fondé sur une série d'automates « à trou » : chaque « trou » correspond à une préposition possible introduisant un argument. Le système renvoie quelques erreurs (soit des groupes de mots qui ne sont pas des prépositions, soit des séquences de mots dues à des rencontres de surface sans pertinence linguistique). L'acquisition est effectuée sur le même corpus journalistique que celui utilisé pour l'acquisition des auxiliaires (section précédente). Après validation des résultats par l'utilisateur, les prépositions alternatives sont ajoutées à la table de données.

3.2.3. Les adjonctions

Volem ne gère que les arguments clé d'un verbe. Cependant, certains compléments prépositionnels, traditionnellement considérés comme des modifieurs, jouent un rôle extrêmement important dans les relations à extraire. Par exemple, un achat selon Volem ne fait pas intervenir de montant alors que le montant est quasiment toujours présent quand on se fonde sur l'analyse en corpus. Celui-ci peut donc être considéré comme un argument clé d'un point de vue sémantique.

L'approche développée par les auteurs de FrameNet est du même type (Fillmore & Baker, 2001). La description des schémas prédicatifs au sein de cette ressource se fonde sur l'étude en corpus des réalisations du verbe. Si un complément intervient fréquemment pour un verbe donné, alors celui-ci sera assimilé à un argument, même s'il est considéré comme un ajout dans la grammaire traditionnelle.

Nous avons alors tenté, en dénombrant ces adjonctions au sein des corpus étudiés, de déterminer quelles adjonctions essentielles pouvaient tenir lieu d'argument, et dans quelle mesure celles-ci pouvaient être repérées par une analyse statistique. La méthode se fonde sur le repérage et le regroupement des compléments circonstanciels (temps, lieu, montant...) pour chaque verbe au sein du corpus. Après dénombrement des adjonctions, nous ne retenons que celles au-dessus d'un seuil de 10% (défini manuellement). Cela signifie que nous ne sélectionnons que celles qui sont apparues dans au moins 10% des phrases contenant un verbe donné. Cette méthode nous a permis de compléter les schémas prédicatifs de plus de 80% des verbes de la ressource.

3.3. Les automates patrons

La dernière étape de l'enrichissement de la ressource a consisté en la création d'automates patrons pour chacune des alternances listée dans Volem. Un automate patron est un automate lexicalement vide, encodant une famille d'alternances ; il est instancié par l'ensemble des verbes correspondant à la famille d'alternances visée. Pour cela, il a fallu dans un premier temps identifier les différentes formes de surface que présentent chacune des alternances de Volem, puis réaliser pour chacune d'elles des graphes permettant de les reconnaître.

Ces automates patrons prennent en entrée la ressource que nous avons créée à partir de Volem traduit sous forme d'une table de contraintes, et produisent en sortie autant d'automates que d'alternances à reconnaître pour chaque verbe. Les automates ainsi créés annotent le texte avec les informations que l'on souhaite extraire. En l'occurrence, pour notre extraction portant sur les rachats d'entreprise, les automates reconnaissent les fragments de textes correspondant à l'acheteur, au vendeur, à l'élément vendu, au montant et à la date.

4. Expériences

4.1. Données d'évaluation

Nous avons choisi de travailler sur une tâche d'extraction préciseé : le rachat d'entreprises. Les expériences ont été principalement menées sur le corpus FUSACQ (`www.fusacq.com`). Ce corpus, de 300 000 octets (environ 35 000 mots), est constitué de dépêches relatant des fusions et des acquisitions d'entreprises. Nous avons également utilisé des corpus à large couverture (journal *Le Monde* par exemple) pour compléter les données, comme nous avons déjà eu l'occasion de le voir. Enfin, le corpus FirstInvest, utilisé par (Poibeau, 2003), a servi de corpus d'entraînement du système.

L'utilisation de plusieurs corpus permet d'évaluer les performances en tenant compte (dans la mesure du possible) du genre textuel. On ne trouve pas les mêmes constructions ni les mêmes expressions suivant que l'on a affaire à un corpus journalistique ou à un site web. Nous verrons dans la discussion que cette hypothèse se vérifie dans notre cas, même si nous avons essentiellement utilisé des corpus de faible taille.

4.2. Résultats

Nous avons mis en place deux protocoles d'évaluation, afin d'isoler les éventuels problèmes ; dans l'un, nous passons les règles d'extraction sur un corpus dans lequel les entités nommées ont été annotées grâce à un outil développé au LIPN (TagEN, `www-lipn.univ-paris13.fr/~poibeau/tagen.html`). Dans l'autre, nous utilisons un corpus dans lequel nous avons annoté toutes les entités nommées à la main. Nous pouvons ainsi procéder d'un côté à une évaluation « en conditions réelles », et de l'autre, nous focaliser sur l'évaluation des schémas prédicatifs, indépendamment des erreurs dues à la mauvaise reconnaissance des entités.

Dans le tableau 1, les "relations" correspondent à des structures grammaticales comportant un verbe et ses arguments participant à un rachat d'entreprise.

	Protocole 1	Protocole 2	Nombre total de relations
Nombre de relations repérées	101	184	285
% de relations	35	64	100

Table 1. Tableau des résultats de l'extraction sur le corpus FUSACQ

Seulement un peu plus de la moitié des entités nommées correspondant à un acheteur potentiel ou à un vendeur potentiel a été annotée. L'annotation des entités nommées n'a pas été menée plus avant, étant donnée qu'elle n'est pas au centre de notre étude. Notre outil permet de repérer (dans un texte dans lequel l'annotation des entités nommées est correcte) 65% des relations d'achat.

35% des relations restent malgré tout non repérées. Ceci provient du fait que la syntaxe de la phrase n'est pas gérée en tant que telle (ce n'était pas le but premier de cette étude). Plus précisément, les constructions suivantes ne sont pas gérées :

- les subordonnées relatives ;
- les verbes introducteurs précédés d'un verbe marquant soit le passé, soit le futur (ambiguïté sémantique possible. Ex. : "Bull vient d'annoncer le rachat de CP8 à Schlumberger") ;
- les structures complexes (ex. :"COMPANY1 s'est diversifié à travers l'acquisition de COMPANY2") ;
- L'alternance passive sans groupe prépositionnel (non encodé dans Volem pour les verbes qui nous intéressent (ex. : "COMPANY a été racheté") ;
- Les structures faisant intervenir un pronom (absence de système de résolution des anaphores).

Verbe	Alternances	nombre d'occurrences de l'alternance (FUSACQ)	nombre d'occurrences de l'alternance (Corpus général)
racheter	caus_2np	37	16
	caus_2np_pp	6	12
	pas_etre_part_np_pp	6	12
revendre	caus_2np	1	0
acheter	caus_refl_pr_np	2	0
	caus_2np	0	16
vendre	pas_etre_part_np_2pp	2	0
	caus_2np	2	0
	caus_2np_pp	2	0
acquérir	caus_2np	44	0
	caus_2np_pp	1	4
	pas_etre_part_np_pp	0	16
céder	caus_2np_pp	24	4
	caus_2np	9	0
	pas_etre_part_np_2pp	2	8
fusionner	aucune occurence	0	0
détenir	caus_2np	5	0
	pas_etre_part_np_pp	1	12
offrir	caus_2np_pp	1	16
	caus_2np	1	0
reprendre	pas_etre_part_np_pp	11	16
	caus_2np	12	8

Table 2. Répartition des alternances selon les verbes dans les phrases extraites des corpus (FUSACQ annoté et extrait du corpus général)

Les adjonctions intéressantes pour une tâche d'extraction sont la date et le montant de la transaction. Les corpus sur lesquels nous avons fait nos expériences ont montré ce point, même s'ils sont de taille trop faible pour donner des chiffres significatifs statistiquement.

Nous obtenons des performances légèrement inférieures que (Poibeau, 2003), essentiellement dû au fait que nous ne gérons pas directement l'interaction entre les schémas prédicatifs et la syntaxe de la phrase. Il a cependant été montré par de nombreux auteurs que l'utilisation de ressources existantes permettaient un gain en temps et en fiabilité par rapport à des ressources *ad hoc* (Giuglea & Moschitti, 2004). La prise en compte de l'analyse phrastique et son interaction avec le niveau prédicatif est un prolongement évident et nécessaire de ce travail, encore préliminaire.

4.3. Discussion

Nous avons vu que Volem est incomplet du fait qu'il ne gère pas la polysémie et que toutes les alternances n'y sont pas codées. Peut-on ajouter automatiquement aux entrées de Volem les alternances que cette ressource n'encode pas ? Plusieurs expériences ont montré qu'il est possible, par des méthodes statistiques, de sélectionner des schémas de sous-catégorisation acceptables pour un verbe donné (Salmon-Alt & Chesley, 2006) (Briscoe & Carroll, 1997). Le typage des arguments du verbe reste toutefois un point nécessaire et peu exploré à l'heure actuelle : il doit être fait manuellement.

Un autre point intéressant est la variation de l'usage du verbe suivant le corpus. On met au jour, même sur des corpus de taille modeste, des variations d'usage : une alternance peut ainsi être employée fortement dans un corpus, beaucoup moins dans un autre *cf tableau2)*. Nous faisons l'hypothèse que ces variations sont liées au domaine et au style d'écriture propre aux différents genres textuels. Ainsi, l'alternance "caus_2np" du verbe "détenir" (*Thales détient 100 % de Raydeon*) constitue 78% des variations syntaxiques pour le verbe "détenir" dans le corpus FirstInvest, et 80% dans FUSACQ, mais n'apparaît quasiment pas dans le corpus tiré de journaux non spécialisés. On voit ici la validité de l'hypothèse sous-jacente dans beaucoup d'études : « un sens par corpus ». En effet, un sens est souvent prégnant, et il ne s'agit pas toujours du sens le plus répandu quand on a affaire à un corpus spécialisé.

5. Conclusion et Perspectives

Les ressources pour le français sont beaucoup moins complètes que celles pour l'anglais. La ressource pour le français qui nous a semblé la plus adaptée à l'extraction d'information (Volem), présente des manques (non gestion de la polysémie, couverture faible, alternances non encodées...) qu'il est cependant possible de combler par des méthodes semi-automatiques. Les résultats obtenus pour une tâche d'extraction pour l'anglais (Giuglea & Moschitti, 2004) montrent que l'extraction à base de ressources à large couverture permet d'obtenir de bons résultats et évite de redévelopper de manière *ad hoc* des connaissances pour chaque nouvelle application. Cet article a cherché à montrer comment compléter Volem, notamment au niveau de la structure argumentale du verbe. Il reste à définir une méthode semi-automatique pour l'ajout des alternances non référencées par Volem et, d'une manière plus générale, à augmenter la couverture de la ressource.

References

BRISCOE T. & CARROLL J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ANLP Conference*, p. 356–363, Washington, USA.

FILLMORE C. & BAKER C. (2001). Frame semantics for text understanding. In *Proceedings of the NAACL workshop « WordNet and Other Lexical Resources Workshop »*, Pittsburgh, USA.

FILLMORE C., JOHNSON C. & PETRUCK M. (2003). Background to framenet. *International Journal of Lexicography*, **16.3**, 235–250.

GARDENT C., GUILLAUME B., FALK I. & PERRIER G. (2005). Maurice Gross' Grammar Lexicon and Natural Language Processing. In *Proceedings of the 2nd Language and Technology Conference*, Poznan, Poland.

GIUGLEA A.-M. & MOSCHITTI A. (2004). Knowledge discovering using framenet, verbnet and propbank. In *International Workshop on Mining for and from the Semantic Web*, Seattle, USA.

GROSS M. (1975). *Méthodes en syntaxe*. Paris: Hermann.

KIPPER-SCHULER K. (2003). Verbnet: a broad coverage, comprehensive, verb lexicon. *Ph.D. Thesis*.

LOWE J., BAKER C. & FILLMORE C. (1997). A frame-semantic approach to semantic annotation. In *Proceedings ANLP workshop: Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, USA.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, **31.1**, 71–106.

PITEL G. (2006). Framenet, théorie, produit, processus, multilingualité et connexions. In *Journée Atala : « Autour de FrameNet et de la Sémantique Lexicale Multilingue »*.

POIBEAU T. (2003). *Extraction automatique d'information, du texte brut au web sémantique*. Paris: Hermes.

SAINT-DIZIER P., FERNANDEZ A., VAZQUEZ G., KAMEL M. & BENAMARA F. (2002). The Volem Project : a Framework for the Construction of Advanced Multilingual Lexicons . In *Language Technology 2002*, p. 123–142: Springer Verlag, Lecture Notes.

SALMON-ALT S. & CHESLEY P. (2006). Automatic extraction of subcategorization frames for french. In *Proceedings of the Language and Resource Evaluation Conference 2006*, Genoa, Italy.