La complémentarité des approches manuelle et automatique en acquisition lexicale

Cédric Messiant [1],Takuya Nakamura[2], Stavroula Voyatzi[2]
(1) Laboratoire d'Informatique de Paris-Nord
CNRS UMR 7030 et Université Paris 13
99, avenue Jean-Baptiste Clément, F-93430 Villetaneuse France
(2) Laboratoire d'Informatique Gaspard-Monge
CNRS UMR 8049 IGM-LabInfo et Université de Marne-la-Vallée
5 Bd Descartes, Champs-sur-Marne
77454 Marne-la-Vallée Cedex 2
cedric.messiant@lipn.univ-paris13.fr, takuya.nakamura@univ-mlv.fr,
voyatzi@univ-mlv.fr

Résumé. Les ressources lexicales sont essentielles pour obtenir des systèmes de traitement des langues performants. Ces ressources peuvent être soit construites à la main, soit acquises automatiquement à partir de gros corpus. Dans cet article, nous montrons la complémentarité de ces deux approches. Pour ce faire, nous utilisons l'exemple de la sous-catégorisation verbale en comparant un lexique acquis par des méthodes automatiques (LexSchem) avec un lexique construit manuellement (Le Lexique-Grammaire). Nous montrons que les informations acquises par ces deux méthodes sont bien distinctes et qu'elles peuvent s'enrichir mutuellement.

Abstract. Lexical resources are essentially created to obtain efficient text-processing systems. These resources can be constructed either manually or automatically from large corpora. In this paper, we show the complementarity of these two types of approaches, comparing an automatically constructed lexicon (LexSchem) to a manually constructed one (Lexique-Grammaire), on examples of verbal subcategorization. The results show that the information retained by these two resources is in fact different and that they can be mutually enhanced.

Mots-clés : verbe, syntaxe, lexique, sous-catégorisation.

Keywords: verb, syntax, lexicon, subcategorization.

1 Introduction

Les applications de traitement automatique des langues (TAL) ont de plus en plus besoin d'informations lexicales. La disponibilité de telles ressources et leur utilisabilité dans les applications de TAL restent pourtant relatives.

Deux approches sont possibles pour la constitution de telles ressources : la construction de lexiques "à la main" et l'acquisition automatique de ressources à partir de corpus. Le plus souvent, ces deux approches sont opposées et il existe à l'heure actuelle peu de collaborations pour étudier les différences dans les informations obtenues et ce que les approches pourraient apporter l'une à l'autre. L'objectif de ce travail est de confronter les approches automatiques aux besoins des linguistes. Pour ce faire, nous utilisons l'exemple de la sous-catégorisation verbale en comparant un lexique acquis par des méthodes automatiques (LexSchem (Messiant *et al.*, 2008)) avec un lexique construit manuellement (Le Lexique-Grammaire (Gross, 1975)). Nous avons observé leurs différences et leurs similitudes pour 20 verbes variés du français.

Après avoir dressé un bref état de l'art des lexiques et des méthodes d'acquisition existantes, nous présentons les deux ressources qui ont été choisies pour cette étude. La section 4 montre que ces deux approches sont complémentaires. La conclusion donne des perspectives pour l'utilisation conjointe de ces approches.

2 État de l'art

La constitution de ressources lexicales a été dans un premier temps manuelle mais des méthodes d'acquisition automatique ont été par la suite explorées.

2.1 Les lexiques existants pour le français

Plusieurs ressources lexicales syntaxiques pour le français ont été développées depuis de nombreuses années. Les objectifs de ces lexiques sont de définir, pour chaque lemme verbal donné, ses différents emplois et, pour chacun de ces emplois, son (ou ses) cadre(s) de sous-catégorisation spécifiant le nombre et le type de ses arguments, et les informations complémentaires qui s'y rapportent.

Le **dictionnaire syntaxique des verbes français** créé par Jean Dubois et Françoise Dubois-Charlier (Dubois & Dubois-Charlier, 1997), mis à la disposition du grand public sur le site internet du laboratoire MoDyCo, est une classification sémantico-syntaxique des verbes manuellement construite par ces deux linguistes, dont les principes sont proches de ceux du LG. On compte dans ce dictionnaire 12 130 verbes et 25 610 entrées (chaque entrée correspond à un couple verbe - schéma de sous-catégorisation).

DicoValence (van den Eynde & Mertens, 2006), successeur du lexique PROTON, est un dictionnaire syntaxique manuellement construit dans le cadre méthodologique de l'Approche Pronominale (van den Eynde & Blanche-Benveniste, 1978). Pour identifier la valence d'un prédicat, i.e. ses dépendants et leurs caractéristiques, l'Approche Pronominale exploite la relation qui existe entre les dépendants dits lexicalisés (réalisés sous forme de syntagmes) et les pronoms qui couvrent « en intention » ces lexicalisations possibles. DicoValence comporte les cadres

de valence de 3 738 verbes, répartis en 8 313 entrées. Nous pouvons également mentionner d'autres ressources comme **LexValf** (Salkoff & Valli, 2005) dont les principes de base sont ceux de grammaire en chaîne, **DiCo-LAF** (Mel´cuk & Polguère, 2006), centré sur la modélisation formelle des collocations et de la dérivation sémantique du français, **DicoLPL** (Vanrullen *et al.*, 2005) ou encore le **Trésor de la Langue Française informatisé** (**TLFI**) (Dendien & Pierrel, 2003).

Des ressources lexicales ont également été acquises semi-automatiquement. C'est le cas notamment de **TreeLex** (Kupść, 2007), acquis automatiquement à partir du corpus arboré de Paris 7 (Abeillé *et al.*, 2003) ou du **Lefff 2** (lexique des formes fléchies du français) (Sagot *et al.*, 2006). **SynLex** (Gardent *et al.*, 2006) est un lexique de sous-catégorisation verbale du français, créé à partir des tables du **LexiqueGrammaire**, et complété manuellement.

2.2 Les méthodes d'acquisition automatique de ressources

Depuis le début des années 90, des chercheurs en traitement automatique des langues ont exploré des méthodes d'acquisition d'informations lexicales à partir de corpus et en particulier de lexiques de schémas de sous-catégorisation (Brent, 1993; Manning, 1993; Briscoe & Carroll, 1997).

Dans un premier temps, ces méthodes ne permettaient d'acquérir qu'un nombre réduit de schémas contenant peu d'informations sur les arguments. La disponibilité de corpus de taille conséquente et les performances des analyseurs syntaxiques ont permis à ces méthodes de devenir de plus en plus efficaces et d'obtenir des informations plus détaillées sur les arguments (Preiss *et al.*, 2007). En effet, la plupart de ces méthodes consistent à acquérir des schémas de sous-catégorisation à partir des sorties d'un analyseur syntaxique et d'en filtrer les erreurs par des méthodes statistiques. La principale difficulté provient alors des erreurs effectuées par l'analyseur syntaxique et de la mise en place du filtrage adéquat.

Dans un premier temps, ces méthodes ne permettaient d'acquérir qu'un nombre réduit de schémas contenant peu d'informations sur les arguments. La disponibilité de corpus de taille conséquente et les performances des analyseurs syntaxiques ont permis à ces méthodes de devenir de plus en plus efficaces et d'obtenir des informations plus détaillées sur les arguments (Korhonen *et al.*, 2000; Preiss *et al.*, 2007). En effet, la plupart de ces méthodes consistent à acquérir des schémas de sous-catégorisation à partir des sorties d'un analyseur syntaxique et d'en filtrer les erreurs par des méthodes statistiques. La principale difficulté provient alors des erreurs effectuées par l'analyseur syntaxique et de la mise en place du filtrage adéquat.

Si la plupart des travaux ont d'abord porté sur l'anglais, ces méthodes ont également été utilisées pour acquérir des informations lexicales pour d'autres langues, comme l'allemand (Schulte im Walde, 2002) ou l'italien (Dino Ienco & Bosco, 2008). Jusqu'en 2006, il y avait peu d'études de ce type pour le français. Le premier travail publié est celui de Paula Chesley et Suzanne Salmon-Alt (Chesley & Salmon-Alt, 2006). Cependant, cette étude ne concernait qu'une centaine de verbes et n'a, à notre connnaissance, pas été poursuivie.

Les ressources acquises à partir de corpus ne sont pas aussi complètes et précises que les ressources construites à la main mais elles apportent des informations qui ne sont souvent pas disponibles dans les travaux manuels comme par exemple la fréquence des schémas de sous-catégorisation en corpus.

3 Une expérience à partir du Lexique-Grammaire et de LexSchem

Afin d'explorer la complémentarité de l'approche manuelle et de l'approche automatique, nous avons travaillé sur deux ressources : le **Lexique-Grammaire** et **LexSchem**.

3.1 Le Lexique-Grammaire

Le lexique syntaxique du **Lexique-Grammaire** [1] (LG) est un dictionnaire syntaxique constitué d'un ensemble de tables (matrices binaires). Chaque table regroupe les éléments prédicatifs (verbes, adjectifs, noms), qui partagent tous la propiété de base d'entrer dans une structure de phrase simple définitoire de la table (construction type). Une phrase simple est définie par le nombre et la nature morpho-syntactique et sémantique des arguments. Chaque table comprend également un ensemble de propriétés distributionnelles, transformationnelles et sémantiques, que vérifient (+), ou non (-), les éléments prédicatifs qui figurent en en-têtes des lignes.

A peu près 6 000 verbes simples graphiquement différents ont été examinés pour le français et donnent lieu à environ 11 000 emplois verbaux simples (ou entrées), dont :
– 3 000 emplois verbaux simples à constructions complétives répartis dans 18 tables,
– 8 000 emplois verbaux simples à constructions non complétives répartis dans 43 tables.
Une sélection des tables du LG du français (60%) est mise à disposition du grand public [2] sous la licence LGPL-LR : en ce qui concerne les verbes simples, on peut compter plus de 7 000 entrées qui proviennent de 35 tables différentes.

3.2 LexSchem

Nous avons mis au point une méthode d'acquisition automatique de schémas de sous-catégorisation à partir de corpus pour le français (Messiant, 2008). Notre système d'acquisition, appelé ASSCI, prend en entrée un corpus analysé par l'analyseur syntaxique **Syntex** (Bourigault *et al.*, 2005) et produit un lexique de sous-catégorisation de verbes.

Une première expérience, menée sur le corpus *LM10* (10 ans du journal *Le Monde*) a permis d'acquérir LexSchem, un lexique de schémas de sous-catégorisation de verbes français (Messiant *et al.*, 2008). Le lexique est composé de plus de 11000 entrées qui correspondent à des couples verbe - schéma. LexSchem concerne plus de 3200 verbes français et près de 300 schémas de sous-catégorisation différents. Le lexique est téléchargeable librement et une interface web permet de le consulter facilement[3]. Pour chacune des entrées, le lexique fournit son nombre d'occurences en corpus, sa fréquence relative (fréquence d'apparition dans le corpus d'un schéma de sous-catégorisation pour un verbe) ainsi que 5 exemples tirés du corpus. Nous avons comparé les entrées de ce lexique pour 25 verbes français choisis pour leur hétérogénéité

[1] Le lexique-grammaire est une théorie et une pratique de la description exhaustive des langues, inspirée de la théorie distributionnelle et transformationnelle de Zellig S. Harris. La description du français a d'abord été développée au LADL par une équipe de linguistes et d'informaticiens dirigée par Maurice Gross depuis la fin des années 1960, et continue d'être maintenue et enrichie par une équipe informatique-linguistique de l'Institut Gaspard-Monge de l'Université Paris-Est Marne-la-Vallée.

[2] `http://infolingu.univ-mlv.fr`

[3] `http://www-lipn.univ-paris13.fr/~messiant/lexschem.html`

avec une ressource référence construite à partir du TLFI [4]. Nous avons choisi le TLFI car il est librement disponible et que son format permettait de rapidement produire une référence pour un nombre limité de verbes [5]. Cette évaluation a donné une précision de 0.83, un rappel de 0.59 et une F-mesure de 0.69. Ces résultats correspondent à ceux de l'état de l'art pour des expériences comparables (par exemple, (Preiss *et al.*, 2007) malgré les différences liées à la langue). Néanmoins, de tels résultats posent la question de l'utilisabilité d'une ressource avec environ 30% d'erreurs. Les erreurs observées dans LexSchem proviennent de sources variées : présence de modifieurs dans le schéma, erreurs d'analyse syntaxique, etc... Certaines de ces erreurs seront détaillées dans la section 4.1.

4 Limites et complémentarité des deux approches

Nous avons comparé les entrées de LexSchem[6] avec celles du LG pour 15 verbes séparés en 3 groupes selon leur productivité dans *LM10* : haute, moyenne et basse fréquence. Nous avons ensuite choisi 5 verbes pour chacun de ces groupes : *savoir*, *demander*, *appeler*, *tourner* et *travailler* pour les verbes à haute fréquence ; *décerner*, *fasciner*, *appuyer*, *rémunérer* et *valider* pour les verbes à fréquence moyenne ; *contrecarrer*, *fluctuer*, *approvisionner*, *insérer* et *cerner* pour les verbes à basse fréquence. Nous avons également ajouté à cette liste cinq verbes susceptibles d'être utilisés dans le domaine économique et boursier (*augmenter*, *baisser*, *chuter*, *clôturer* et *fluctuer*).

Cette comparaison nous a permis de mettre en évidence les limites des deux approches. Elle nous a aussi montré pourquoi ces approches sont complémentaires et comment elles peuvent s'enrichir mutuellement.

4.1 Les limites de l'acquisition à partir de corpus et de LexSchem

Malgré les progrès récents, l'acquisition de ressources lexicales à partir de corpus produit un grand nombre d'erreurs (voir section 3.2). La confrontation des lexiques extraits par ces méthodes à des ressources construites manuellement permet de mettre en évidence les sources des erreurs et les lacunes de l'approche automatique.

De nombreuses erreurs d'étiquetage morpho-syntaxique ou d'analyse syntaxique sont répercutées dans le lexique. Pour ce qui concerne les erreurs d'étiquetage, il s'agit dans la plupart des cas de noms étiquetés verbe. Par exemple, dans le groupe nominal *le programme d'armement*, le mot *programme* est étiqueté verbe et ASSCI produit alors le schéma de sous-catégorisation `SUJ:SN_P-OBJ:SP[de+SN]`. Autre exemple, *le CERNA (Centre d'économie industrielle et de finance de MINES ParisTech)* apparaît dans les exemples pour le verbe *cerner*. Notons que ces exemples sont suffisamment fréquents pour produire des schémas de sous-catégorisation incorrects.

[4] `http://atilf.atilf.fr/tlf.htm`

[5] Une étude sur 25 verbes n'est pas suffisante. C'est pourquoi nous travaillons actuellement à l'évaluation de LexSchem sur environ 1500 verbes français en utilisant TreeLex comme ressource référence.

[6] La version de LexSchem utilisée pour les expériences de cette section n'est pas celle qui est téléchargeable et consultable à l'adresse donnée à la section 3.2. Les améliorations de cette nouvelle version comprennent notamment l'intégration des fonctions syntaxiques aux schémas de sous-catégorisation et le traitement de cas jusqu'alors mal traités.

Par ailleurs, de nombreuses erreurs proviennent de l'analyse syntaxique. Par exemple, les pronominalisations sont souvent mal traitées par l'analyseur comme dans *Leurs obsessions nous cernent* qui produit le schéma `SUJ:SN` (c'est-à-dire une construction de type intransitive). Il existe plus généralement de nombreux compléments d'objets non rattachés, notamment quand des incises brouillent l'analyse. Ce problème est très fréquent et on pourrait imaginer un traitement *a posteriori* pour tenter de rattacher ces compléments au verbe. Cependant, un tel post-traitement poserait alors la question de la "généricité" de la méthode et de son indépendance par rapport à l'analyseur syntaxique utilisé.

Certains phénomènes linguistiques sont assez difficiles à prendre en compte par un système d'acquisition qui se fonde sur des indices de surface. C'est le cas notamment des adverbes (qui sont souvent des modifieurs mais qui sont susceptibles d'être inclus dans la structure argumentale comme dans *Le couteau coupe bien*) ou les propositions interrogatives indirectes. En effet, comment différencier *Il se demande qui est passé en premier.* de *Le président nomme qui est le plus méritant.* de manière automatique ?

La prise en compte de ces phénomènes par un système d'acquisition automatique est possible mais nécessiterait d'introduire la notion de classe sémantique sur le verbe (verbe de type interrogation) pour acquérir les bonnes informations. Cet exemple montre bien les limites de l'acquisition automatique : lorsqu'on connaît les verbes concernés par un phénomène précis, n'est-il pas préférable de corriger la ressource *a posteriori* plutôt que d'ajouter des informations *a priori* ?

Certains schémas de sous-catégorisation sont valides mais nécessiteraient d'être accompagnés d'informations sur les restrictions de sélection. Par exemple, le verbe *tourner* accepte un emploi absolu (sans complément) comme dans *Le vent tourne* ou *Le lait tourne* mais cet emploi impose des restrictions quant à la sémantique du nom en position sujet. Même si les principales têtes nominales apparaissant dans les différentes positions argumentales sont conservées par ASSCI, aucun calcul n'est effectué à l'heure actuelle pour repérer les restrictions de sélection (on envisage de donner une approximation de ce degré de figement à partir des informations enregistrées par le système). Dans le LG, chacune de ces expressions est considérée comme figée et est répertoriée dans une table qui lui est propre (voir section 4.2).

Les informations disponibles dans un lexique acquis à partir de corpus ne sont pas aussi riches que celles qu'on trouve dans une ressource comme le LG. En particulier, LexSchem ne fournit aucune information sur la sémantique des arguments du verbe dans les schémas de sous-catégorisation. Par exemple, dans le LG, le verbe *savoir* a un schéma `N0hum V Vinf` (le sujet du verbe doit dans ce cas avoir le trait "humain"). Ce type d'information est totalement absent d'un lexique acquis automatiquement comme LexSchem et il semble pour l'instant difficile de l'ajouter sans intervention humaine. Toutefois, ce type d'information n'est pas sans poser de problème dans le cas de métonymies ou de métaphores.

4.2 Les limites des approches manuelles et du Lexique-Grammaire

La création des données syntaxiques du LG est entièrement manuelle. Indépendamment de la question de coûts humains et temporels, cette méthode recèle quelques problèmes. Nous en discutons quelques uns :

– Il est relativement difficile de maintenir la cohérence de classification : par souci d'exhaustivité, la stratégie de classification était de conserver, plutôt que d'exclure. Ainsi, un même verbe risque de se retrouver dans deux classes différentes. Par exemple, le verbe savoir qui

sélectionne une complétive directe possède deux emplois : l'un classé dans la construction `N0hum V Que P =` : *Luc sait que Léa est à Paris*, et l'autre classé dans la table 10, entrant dans la construction `N0hum V par N2hum Que P =` : *Luc a su par Max que Léa est à Paris*. S'agit-il vraiment d'homonymie ? Nous ne pouvons pas entrer dans le détail mais en tout cas, pour l'acquisition automatique de valence, les deux emplois sont regroupés sous un seul.

– La distinction en plusieurs emplois doit être minutieusement contrôlée, sinon elle créera des homonymes à profusion. Les critères essentiels sont, par définition, formels, de telle sorte que l'ensemble de propriétés vérifiées par un verbe donné doit être clairement distinct de celui de son (ou ses) homonyme(s), si le verbe possède plusieurs emplois. Ce critère peut être difficile à respecter s'il s'agit d'un simple verbe bi-valent, par exemple. Ainsi, pour le verbe très courant comme *travailler*, le LG dispose de 15 emplois différents dont quatre appartiennent à la classe appelée 32R3 (classe de constructions transitives résiduelles). Les quatre emplois du verbe travailler classés dans 32R3 sont tous distingués par différents objets directs lexicaux :

  (1) *Max travaille la balle*
  (2) *Max travaille son texte*
  (3) *Max travaille cette discipline*
  (4) *Max travaille l'opinion publique*

  Le sens du verbe *travailler* est certes différent dans chaque cas, mais les critères syntaxiques qui les différencient sont minces. Ce sont des exemples à la frontière des expressions libres-expressions figées. Á propos de ce verbe, LexSchem donne simplement deux sous-catégorisations possibles : `SUJ:SN_OBJ:SN` et `SUJ:SN`. Il manque totalement d'exemple transitif simple. Cela étant, on peut se poser la question légitime de savoir si cette "finesse" de distinction en plusieurs emplois d'une construction transitive est nécessaire pour les besoins du TAL ?

– Par ailleurs, les classes dites "résiduelles", marquées par R comme 32R3, ont tendance à regrouper tous les exemples qui n'obéissent pas strictement à des critères de classification. Elles servent un peu de "fourre-tout" du LG. Généralement les exemples accumulés dans cette classe attendent d'être reclassés dans d'autres classes. Seule l'amélioration continue du LG allégera ce problème.

4.3 La complémentarité des méthodes

Le plus souvent, les ressources construites à la main ne comprennent aucune information sur la fréquence des schémas de sous-catégorisation. Il est possible d'utiliser des méthodes d'acquisition automatique pour acquérir ces informations statistiques sur la langue et inclure ces informations dans des ressources manuelles.

Lors de la comparaison de LexSchem et du LG, nous avons remarqué que les schémas de sous-catégorisation recensés dans les deux ressources ne sont pas toujours les mêmes. Dans la plupart des cas, ces schémas sont présents dans le LG et absents de LexSchem. Par exemple, dans LexSchem, le verbe *valider* est assorti de trois schémas de sous-catégorisation, dont deux correspondent aux constructions respectivement du passif `SUJ:SN_P-OBJ:SP[par+SN]`, et du participe passé supporté par le verbe être `SUJ:SN`. LexSchem, ne pouvant pas différencier de manière automatique ce type de variantes transformationnelles des variantes formelles significatives pour la valence verbale, il produit systématiquement ces deux cadres de sous-

catégorisation pour tous les verbes vérifiant ces constructions dans le corpus. En revanche, dans le LG, le verbe *valider* est représenté dans deux classes différentes qui sont définies par les constructions type :

– Table 6 `N0 V N1`, avec `N1 = : Qu P + si P + N`
= : *L'entreprise va valider (que vos compétences sont adaptées à son projet + si vos compétences sont adaptées à son projet + vos compétences)*
– Table 32RA `N0 V N1 (E + de N2)`, avec `V = : rendre (E + plus) adjectif`
= : *Le tuteur a validé le certificat de stage d'un tampon*

Néanmoins, il arrive également qu'on détecte un schéma de sous-catégorisation valide (selon les exemples extraits du corpus) dans LexSchem mais absent du LG. Par exemple, dans LexSchem, le verbe *tourner* produit quatre schémas de sous-catégorisation, dont le schéma `SUJ:SN_P-OBJ:SP[autour de+SN]` = : *Le déficit public tournera autour des 2% du PIB.* (le verbe *tourner* a ici le sens de *se dérouler*, *évoluer*). Mais, cette construction est absente du LG. L'utilisation conjointe des deux approches permettrait donc d'obtenir un lexique plus complet.

Des études pour l'anglais ont montré que les schémas de sous-catégorisation d'un verbe pouvaient varier lorsque ce verbe est employé dans une langue de spécialité (Wattarujeekrit *et al.*, 2004). En observant le verbe *fluctuer*, nous avons remarqué que certains schémas de sous-catégorisation étaient présents dans LexSchem mais pas dans le LG. Ce schéma est absent du LG en raison de sa nature spécifique au domaine boursier. Une expérience sur cinq verbes utilisés le plus souvent dans un cadre économique ou boursier dans le corpus LM10 (*augmenter*, *baisser*, *chuter*, *clôturer* et *fluctuer*) nous a permis de vérifier cette hypothèse. Étant donné la nature et l'objectif de cette tâche, nous avons modifié les paramètres d'ASSCI afin qu'il soit plus "permissif" et qu'il fournisse un plus grand nombre de schémas de sous-catégorisation pour ces verbes. Ainsi, nous avons pu observer dans quelles proportions l'étude à partir de corpus diffère de la construction manuelle de ressources dans le cadre d'un domaine précis de la langue.

Les verbes économiques examinés ont ceci de particulier qu'ils expriment à peu près tous la variation en quantité ou en nombre. Les compléments prépositionnels qui les accompagnent ne sont pas des compléments essentiels au sens strict du terme : pratiquement tous sont des adverbiaux. Ils échappent, par conséquent, à la grille de compléments essentiels qui définit une structure de phrase du LG. Par conséquent, la méthode statistique qui puisse les détecter est complémentaire de la méthode manuelle, du point de vue de leur détection. Il n'en reste pas moins qu'il est difficile de les valider d'emblée comme compléments sous-catégorisés pour plusieurs raisons. Ils réalisent souvent des arguments sémantiques et la correspondance entre forme et notion sémantique n'est pas bi-univoque : à un argument peut correspondre plusieurs formes. On peut énumérer, à propos des verbes examinés, des arguments sémantiques comme "mesure relative", "mesure absolue", etc. Par exemple, un verbe foncièrement intransitif comme *augmenter* est souvent accompagné d'un complément de mesure relative adverbiale, mais ce dernier peut prendre plusieurs formes, à part la réalisation canonique sous forme de `de DetNum\%` :

(1) *La valeur X a augmenté de 5%.*

(2) *La valeur X augmente à un rythme annuel de 4%.*

de 5% et *à un rythme annuel de 4%* sont, plus ou moins, des variantes formelles de l'argument

de mesure relative, la différence de formes n'est pas foncièrement significative du point de vue de la valence du verbe augmenter. Pour le LexSchem, ils sont deux sous-catégorisations différentes, cependant. L'intervention manuelle pour regrouper ce genre de variantes formelles semble nécessaire.

Cette étude devra être approfondie, notamment par l'acquisition de schémas de sous-catégorisation à partir d'un corpus de spécialité comme le droit ou la médecine. Cette ressource pourra fournir une base de travail à des linguistes qui pourront la valider, l'enrichir et l'affiner.

5 Conclusion

Nous avons montré que la construction manuelle de ressources lexicales et l'acquisition automatique de ressources sont deux approches complémentaires. Plutôt que d'opposer ces approches, il peut être bénéfique de travailler à un rapprochement des approches. Les ressources manuelles peuvent servir à valider les lexiques acquis automatiquement et à leur apporter des informations qu'il est difficile d'acquérir à partir de corpus (par exemple, le rôle sémantique des arguments, ...). Les ressources acquises à partir de corpus peuvent constituer une base de travail pour la constitution de ressources pour un domaine de spécialité ou pour enrichir des lexiques construits à la main avec des informations statistiques par exemple.

A la suite de cette étude, les perspectives sont multiples : définition d'une méthode pour faciliter l'utilisation conjointe des deux approches, par exemple par la mise en place d'une "interface d'enrichissement de ressource" ou harmonisation des formats utilisés. Étant donné les différences notables observées entre LexSchem et le Lexique-Grammaire pour des verbes du domaine économique ou boursier, nous envisagons d'acquérir automatiquement une ressource à partir d'un corpus de spécialité.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In A. ABEILLÉ, Ed., *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Kluwer Academic Publishers : Dordrecht.

BOURIGAULT D., JACQUES M.-P., FABRE C., FRÉROT C. & OZDOWSKA S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan.

BRENT M. R. (1993). From grammar to lexicon : Unsupervised learning of lexical syntax. *Computational Linguistics*, **19**, 203–222.

BRISCOE T. & CARROLL J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, p. 356–363, Washington, DC.

CHESLEY P. & SALMON-ALT S. (2006). Automatic extraction of subcategorization frames for french. In *Language Resources and Evaluation Conference (LREC)*, Genua (Italy).

DENDIEN J. & PIERREL J.-M. (2003). Le trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues*, **44 (2)**.

DINO IENCO S. V. & BOSCO C. (2008). Automatic extraction of subcategorization frames for italian. In E. L. R. A. (ELRA), Ed., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les verbes français*. Paris : Larousse.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2006). Extraction d'information de sous-catégorisation à partir des tables du ladl. In *Actes de Traitement Automatique des Langues Naturelles*, Louvain, Belgique.

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

KORHONEN A., GORRELL G. & MCCARTHY D. (2000). Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

KUPŚĆ A. (2007). Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. In *TALN 2007*, Toulouse.

MANNING C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Meeting of the Association for Computational Linguistics*, p. 235–242.

MEL´CUK I. & POLGUÈRE A. (2006). Dérivations sémantiques et collocations dans le DiCo/LAF. *Langue française*, **150**, 66–83.

MESSIANT C. (2008). A subcategorization acquisition system for French verbs. In *Proceedings of the ACL-08 : HLT Student Research Workshop*, p. 55–60, Columbus, Ohio : Association for Computational Linguistics.

MESSIANT C., KORHONEN A. & POIBEAU T. (2008). Lexschem : A large subcategorization lexicon for french verbs. In *Language Resources and Evaluation Conference (LREC)*, Marrakech.

PREISS J., BRISCOE T. & KORHONEN A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Meeting of the Association for Computational Linguistics*, p. 912–918, Prague.

SAGOT B., CLÉMENT L., DE LA CLERGERIE E. & BOULLIER P. (2006). The Lefff 2 syntactic lexicon for French : architecture, acquisition, use. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Genua (Italy).

SALKOFF M. & VALLI A. (2005). A dictionary of french verbal complementation. In *Actes de Language and Technology Conference. Human Language and Technologies as a Challenge for Computer Science and Linguistics. In memory of M. Gross and A. Zampolli*, Poznan, Poland.

SCHULTE IM WALDE S. (2002). A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, p. 1351–1357, Las Palmas de Gran Canaria, Spain.

VAN DEN EYNDE K. & BLANCHE-BENVENISTE C. (1978). Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale. *Cahiers de Lexicologie*, **32**, 3–27.

VAN DEN EYNDE K. & MERTENS P. (2006). *Le dictionnaire de valence Dicovalence : manuel d'utilisation*. Leuven : Manuscript.

VANRULLEN T., BLACHE P., PORTES C., RAUZY S., MAEYHIEUX J.-F., GUÉNOT M.-L., JEAN-MARIE-BALFOURIER & BELLENGIER E. (2005). Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan.

WATTARUJEEKRIT T., SHAH P. K. & COLLIER N. (2004). Pasbio : predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, **5**.