HAL
open science

Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré

Anna Kupść

To cite this version:

Anna Kupść. Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. TALN 2007, 2007, Toulouse, France. halshs-00751131

HAL Id: halshs-00751131

https://shs.hal.science/halshs-00751131v1

Submitted on 18 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré

Anna Kupść
Université Paris3 / LLF, UMR 7110
et Académie Polonaise des Sciences
akupsc@univ-paris3.fr

Résumé. Nous présentons une expérience d'extraction automatique des cadres de sous-catégorisation pour 1362 verbes français. Nous exploitons un corpus journalistique richement annoté de 10 000 phrases dont nous extrayons 12 510 occurrences verbales. Nous évaluons dans un premier temps l'extraction des cadres basée sur la fonction des arguments, ce qui nous fournit 39 cadres différents avec une moyenne de 1.54 cadres par lemme. Ensuite, nous adoptons une approche mixte (fonction et catégorie syntaxique) qui nous fournit dans un premier temps 925 cadres différents, avec une moyenne de 3.44 cadres par lemme. Plusieurs méthodes de factorisation, neutralisant en particulier les variantes de réalisation avec le passif ou les pronoms clitiques, sont ensuite appliquées et nous permettent d'aboutir à 235 cadres différents avec une moyenne de 1.94 cadres par verbe. Nous comparons brièvement nos résultats avec les travaux existants pour le français et pour l'anglais.

Abstract. We present our work on automatic extraction of subcategorisation frames for 1362 French verbs. We use a treebank of 10000 sentences from which we extract 12510 verb occurrences. We evaluate the results based on a functional representation of frames and we acquire 39 different frames, 1.54 per lemma on average. Then, we adopt a mixed representation (functions and categories), which leads to 925 different frames, 3.44 frames on average. We investigate several methods to reduce the ambiguity (e.g., neutralisation of passive forms or clitic arguments), which allows us to arrive at 235 frames, with 1.94 frames per lemma on average. We present a brief comparison with the existing work on French and English.

Mots-clés : français, corpus arboré, sous-catégorisation verbale, lexique-grammaire.

Keywords: French, treebank, verbal subcategorization, lexicon grammar.

1 Introduction

Cet article présente une expérience d'extraction de sous-catégorisations verbales pour le français à partir d'un corpus journalistique richement annoté (le corpus arboré de Paris 7).

Un lexique syntaxique est une ressource qui contient l'information sur le potentiel combinatoire d'un prédicat (ex., le verbe *dormir* régit un seul argument, le sujet), mais aussi sur le type de ses arguments (ex., l'adjectif *fier* se combine avec un syntagme prépositionnel en *de*). Ces informations varient d'une langue à l'autre, elles sont donc essentielles pour l'apprentissage et

l'acquisition des langues. Pour le traitement automatique des langues (TAL), les informations sur la structure prédicative sont importantes dans la plupart des applications. (Briscoe & Carroll, 1993) estiment qu'environ la moitié des erreurs des analyseurs syntaxiques repose sur des informations insuffisantes concernant la structure argumentale, tandis que (Carroll & Fang, 2004) montrent une amélioration significative de la performance d'un parseur enrichi avec un tel lexique. Elles jouent également un rôle essentiel pour la génération automatique (Danlos, 1985), la traduction automatique (Han *et al.*, 2000), ou l'extraction d'information, cf. (Surdeanu *et al.*, 2003).

Néanmoins, ce type d'informations est toujours difficilement disponible. Traditionnellement, de telles ressources ont été développées par des experts humains, par ex., (Procter, 1978; Hornby, 1989) (pour l'anglais) ou les lexiques-grammaires du LADL (Gross, 1975; Guillet & Leclère, 1992) et le Dictionnaire explicatif et Combinatoire (DECFC) de (Mel'cuk *et al.*, 1984 1988 1992 1999) (pour le français), ce qui garantit leur bonne qualité, mais elles ne sont pas directement adaptées au traitement automatique. Par contre, les ressources informatisées développées automatiquement en vue des applications TAL, par ex. : (Bourigault & Frérot, 2005; Sagot *et al.*, 2006) (pour le français), utilisent des méthodes statistiques, ce qui rend les résultats moins fiables.

Dans cet article, nous nous basons sur le corpus arboré de Paris7, cf. (Abeillé *et al.*, 2003; Abeillé & Barrier, 2004), pour obtenir les cadres de sous-catégorisation verbales pour le français. À notre connaissance, ce corpus n'a pas encore été exploité pour l'extraction de telles ressources lexicales. Le corpus arboré de Paris7 est un ensemble de textes journalistiques du Journal Le Monde (1989-1993), annotés aux niveaux morphologique et syntaxique, pour les constituants majeurs mais aussi pour les fonctions grammaticales. L'étiquetage a été validé par des experts humains, ce qui fait du corpus une ressource précieuse pour des recherches linguistiques mais aussi pour le développement d'outils de TAL. Pour l'acquisition des cadres de sous-catégorisation verbale, nous nous sommes basés sur une sous-partie du corpus qui comprend des annotations fonctionnelles, soit environ 10 000 phrases (environ 200 000 mots).

L'objectif de ce travail est d'obtenir une liste de cadres de sous-catégorisation utilisables par différents types de grammaires électroniques, ainsi qu'un lexique informatisé, fiable et de haute qualité, qui pourra servir en particulier pour l'évaluation d'autres lexiques syntaxiques obtenus automatiquement. Il s'agit également d'estimer l'ambiguïté des cadres de sous-catégorisation verbale (combien de cadres par verbe ?) et de rechercher les méthodes pour la réduire. Ceci nous permettra de préparer une ressource bien adaptée pour différentes applications TAL.

2 État de l'art

Les travaux sur l'acquisition du lexique syntaxique à partir de treebanks sont relativement nombreux, et la plupart part de données moins riches que les notres. Pour l'anglais, le lexique syntaxique le plus important extrait à partir du corpus arboré (Penn-II Treebank) a été obtenu par (O'Donovan *et al.*, 2004), comme une ressource supplémentaire de l'induction des grammaires lexicalisées. La même technique a été adoptée pour d'autres langues comme l'allemand et le chinois, inter alia. (Sarkar & Zeman, 2000) présentent les résultats d'apprentissage automatique de cadres de sous-catégorisation à partir du corpus arboré du tchèque (Prague Dependency Treebank), tandis que (Marinov, 2004) applique les mêmes techniques sur un treebank bulgare (BullTreebank). Tous ces lexiques sont obtenus à partir des annotations syntagmatiques, i.e., les

fonctions grammaticales sont assignées automatiquement en utilisant des méthodes statistiques. Notre tâche est différente car nous bénéficions des annotations fonctionnelles déjà existantes dans le treebank, sans intermédiaire probabiliste.

Pour le français, les efforts récents pour construire automatiquement des lexiques syntaxiques se sont basés sur les méthodes probabilistes, cf. (Bourigault & Frérot, 2005), (Chesley & Salmon-Alt, 2005). Elles utilisent des corpus qui n'ont que des informations catégorielles, et non fonctionnelles (ce qui pose problème pour distinguer arguments et ajouts). Un autre pôle est répresenté par les travaux sur l'informatisation et l'actualisation du lexique–grammaire de LADL, effectué par (Gardent *et al.*, 2006). C'est une ressource tout à fait précieuse (plus de 8000 lemmes) mais qui n'a pas à notre connaissance été entièrement informatisée ni surtout validée sur corpus.[1] Un autre lexique syntaxique de la large couverture (plus de 3700 verbes), Proton[2] développé selon l'approche pronominale, n'est pas directement utilisable dans les application TAL.

3 Extraction des cadres de sous-catégorisation verbale

3.1 Choix d'informations à extraire

La représentation des cadres de sous-catégorisation se fait différemment selon différents approches : certains modèles théoriques, comme LFG (grammaire lexicale fonctionnelle) privilégient une notation basée sur les fonctions (1a), d'autres comme le LADL privilégient une notation basée sur les catégories (1b), d'autres enfin, comme en HPSG (grammaire syntagmatique guidée par les têtes), ont une approche mixte (1c) :

(1) *laver* :

a. <SUJ, OBJ>

b. N0 V N1

c. <SUJ :NP, OBJ :NP>

Les deux premières représentations ne sont pas complètes parce que les fonctions et les catégories peuvent avoir plusieurs réalisations (par exemple le sujet peut être nominal ou phrastique, tandis qu'un NP postverbal peut être objet direct ou attribut). Comme nous disposons d'un corpus annoté et pour les catégories et pour les fonctions, nous adoptons une approche mixte pour obtenir l'information plus riche. La liste des fonctions et des catégories dans le corpus est indiquée dans le tab. 1. Nous ignorons la fonction MOD qui correspond toujours à des éléments non sous-catégorisés (des modifieurs). Parmi les réalisations possibles des fonctions, nous ignorons les cas avec COORD, puisque la coordination double est très rare.

Pour les compléments prépositionnels (P-OBJ), nous marquons le type de la préposition régie par le verbe. Ceci nous permettra de normaliser les cadres par rapport aux formes passives et actives. Le noyau verbal, VN, contient le verbe principal mais aussi des auxiliaires, éléments négatifs, et les clitiques pronominaux. Selon une suggestion de (Abeillé & Barrier, 2004), nous considérons que le dernier V est la tête sémantique du VN. Les clitiques ont également une fonction indiquée (au niveau du VN) quand ils correspondent aux arguments du verbe.

[1] Le DECFC de Montréal est également en cours d'informatisation mais ne comprend que 514 vocables, qui ne sont pas tous des verbes (avec des informations sémantiques et non seulement syntaxiques).

[2] `http ://bach.arts.kuleuven.be/PA/proton.html`

SUJ	NP, VPinf, Ssub, COORD, VN
OBJ	NP, AP, VPinf, COORD, VN, Sint, Ssub
DE-OBJ	VPinf, PP, Ssub, VN, COORD
A-OBJ	VPinf, PP, VN, COORD
P-OBJ	PP, AdP, COORD, VN, NP
ATO	Srel, PP, AP, NP, VPpart, COORD, VPinf, Ssub
ATS	NP, PP, AP, AdP, VPinf, Ssub, COORD, VPpart, Sint, VN

FIG. 1 – Liste des catégories possibles pour chaque fonction sous-catégorisée. Fonctions : SUJ (sujet), OBJ (objet direct), DE-OBJ (objet indirect en *de*), A-OBJ (objet indirect en *à*), P-OBJ (complément avec une autre préposition), ATO (attribut de l'objet), ATS (attribut du sujet)

3.2 Description de l'expérience

L'extraction de cadres de sous-catégorisation verbale est plus difficile pour le français que pour l'anglais, d'une part à cause des alternances de variantes avec les pronoms clitiques, d'autre part à cause d'un ordre des mots plus libre (un SN (NP) postverbal peut être un sujet inversé par exemple). Nous avons extrait les lemmes des phrases principales du corpus arboré annoté pour les fonctions. Les fonctions sont traitées comme des attributs des syntagmes et non comme des relations entre la tête et les syntagmes. Elles sont notées soit sur les syntagmes de même niveau que le VN (pour les dépendants du verbe) soit sur le VN lui-même (pour les pronoms clitiques). Puisque le VN contient aussi les auxiliaires, nous traitons le dernier verbe dans VN comme le verbe principal, tandis que les auxiliaires sont stockés afin de normaliser les cadres par rapport aux formes passives et actives.

Comme point de départ, nous avons utilisé les cadres extraits directement du corpus, sans aucune modification, et ensuite nous avons fait plusieurs tests pour compacter les cadres.

D'abord, nous avons dégroupé les fonctions accumulées par les clitiques dans le VN. S'il y a plusieurs clitiques attachés au verbe (ex. : le sujet et l'objet dans *Il l'a vue*), leur fonctions sont groupées dans un seul tag (SUJ/OBJ). Il faut donc les séparer. Les cas où un clitique apparaît sans fonction sont normalement ceux qui correspondent à des réfléchis figés (comme pour *s'évanouir*). Nous les conservons comme tels dans nos cadres. Enfin, on peut avoir dans la même phrase un clitique et un argument nominal de même fonction. Ainsi dans une phrase comme : *Paul en mange-t-il beaucoup ?*, on a deux sujets (*Paul* et *il*) et deux objets (*en* et *beaucoup*). Il faut donc éliminer les duplicats des fonctions dans les cadres. Finalement, il y a des cadres qui n'ont pas de sujet spécifié. C'est le cas pour les formes verbales à l'impératif, et nous avons complété leurs cadres avec SUJ. Seules deux lemmes apparaissent toujours sans sujet : il s'agit de *voici* et *voilà* qui sont analysés comme des formes verbales à l'indicatif, et qui ont donc des cadres spécifiques sans sujet.

Nous avons normalisé les cadres par rapport aux formes du passif. On a utilisé une liste de 62 verbes qui sont conjugués avec *être* pour distinguer les formes du passé et du passif. Ainsi si le verbe apparaît dans le corpus avec l'auxiliaire *être* mais qu'il se conjugue avec *avoir*, son cadre est considéré comme passif et transformé en forme active. On ajoute OBJ, tandis que le complément d'agent (s'il est présent), i.e., P-OBJ introduit par la préposition *par* ou *de*, est supprimé (le SUJ est déjà présent dans les deux cadres passifs), et on change ATS en ATO s'il y a un attribut.

Un deuxième type de normalisation concerne les arguments clitiques. Comme nous extrayons les catégories des arguments, nous obtenons dans un premier temps, un cadre différent pour une occurrence avec sujet clitique et une occurrence avec sujet nominal alors que c'est la même sous-catégorisation. Nous avons donc regroupé les résultats.

Nous avons aussi commencé la factorisation des compléments optionnels. Par exemple, si un même verbe a deux cadres SUJ et SUJ OBJ, nous considérons que l'objet est optionnel. On peut donc lui assigner le cadre SUJ (OBJ).

Certaines difficultés viennent des choix d'annotation du corpus. Par exemple, les syntagmes adverbiaux sous-catégorisés ont une fonction syntaxique associée mais pas les adverbes seuls. Donc l'adverbe *bien* n'est pas reconnu comme le complément dans la phrase *Elle va bien*. Les annotations du corpus ne sont faites que pour les arguments qui appartiennent au cadre du verbe de même niveau. Donc on va rater des cas de dépendance à distance comme : *Que peut faire le gouvernement ?* (puisqu'on va extraire deux OBJ pour *peut* et aucun pour *faire*). Tels cas sont cependant assez rares.

3.3 Résultats

Dans cette expérience nous avons utilisé uniquement les verbes dans les phrases principales, soit 1362 verbe lemmes (12510 tokens). Nous comparons une approche uniquement fonctionnelle de la sous-catégorisation (comme (1a)) et une approche mixte qui tient compte également des catégories (comme (1c)).

3.3.1 Extraction de la Sous-catégorisation fonctionnelle

Nous avons normalisé les cadres par rapport au passif et nous n'avons pas utilisé les catégories. Nous ne tenons pas compte de l'ordre des mots, c'est-à-dire que nous considérons un seul cadre pour *Jean pense à Marie* et *À qui pense Jean* que nous notons 'A-OBJ, SUJ' (avec les fonctions en ordre alphabétique). Si l'on tient compte des clitiques réfléchis, qui peuvent être figés, on aboutit à 39 cadres différents avec une moyenne de 1.75 cadres par verbe. Le lemme avec le plus des cadres, 18, est le verbe *être*. Plus de la moitié des verbes (63.3% des lemmes, soit 862 lemmes différents) sont non ambigus et ont un seul cadre.

On peut réduire le nombre des cadres en éliminant le clitique réfléchi pour les verbes qui ont un OBJ ou un A-OBJ correspondant. Le nombre de cadres différents au total (39) comme le maximum de cadres par lemme (18 pour *être*) ne changent pas, mais on réduit la moyenne de cadres par lemme à 1.68. Avec ceci nous arrivons à 65.1% des verbes (888 lemmes) à un seul cadre.

Ensuite, si l'on factorise les arguments optionnels, par exemple les cadres SUJ et SUJ-OBJ, on obtient plus de cadres différents possibles et moins de cadres différents pour chaque verbe. Pour l'objet optionnel, nous avons un cadre de plus, c'est-à-dire 40 cadres en général, avec une moyenne de 1.54 cadres par lemme, tandis qu'il y a 4 verbes à 10 ou plus cadres, qui sont effectivement parmi les plus ambigus du français (*être*, *passer*, *avoir*, *rendre*). Les résultats sont présentés dans le tableau 2. Les cadres pour les 4 verbes avec le plus de cadres sont dans le tableau 3.

Les cadres les plus fréquents sont ceux des verbes à un complément, tout d'abord ceux à objet

	# cadres	moyenne	max. cadres	verbes à un cadre	
				%	#
avec réfléchi	39	1.75	18 (*être*)	63.3%	862
sans réfléchi	39	1.68	18 (*être*)	65.1%	888
SUJ (OBJ)	40	1.54	17 (*être*)	68.9%	939

FIG. 2 – Résultats pour les cadres fonctionnels

```
être (17):  (OBJ), SUJ| A-OBJ, ATS, OBJ, SUJ| A-OBJ, ATS, SUJ| A-OBJ,
  OBJ, SUJ| A-OBJ, SUJ| ATS, DE-OBJ, OBJ, SUJ| ATS, DE-OBJ, SUJ| ATS,
  DE-OBJ, SUJ, refl| ATS, OBJ, P-OBJ, SUJ| ATS, OBJ, SUJ| ATS, P-OBJ,
  SUJ| ATS, SUJ| ATS, SUJ, refl| DE-OBJ, OBJ, SUJ| DE-OBJ, SUJ| OBJ,
  P-OBJ, SUJ| P-OBJ, SUJ
avoir (11):  (OBJ), SUJ| A-OBJ, DE-OBJ, OBJ, SUJ| A-OBJ, OBJ, SUJ|
  A-OBJ, SUJ| ATO, OBJ, SUJ| ATS, OBJ, SUJ| DE-OBJ, OBJ, P-OBJ, SUJ|
  DE-OBJ, OBJ, SUJ| DE-OBJ, SUJ| OBJ, P-OBJ, SUJ| P-OBJ, SUJ
passer (10):  (OBJ), SUJ| A-OBJ, DE-OBJ, SUJ| A-OBJ, OBJ, SUJ| A-OBJ,
  SUJ| ATS, SUJ| DE-OBJ, SUJ| DE-OBJ, SUJ, refl| OBJ, P-OBJ, SUJ|
  P-OBJ, SUJ| P-OBJ, SUJ, refl
rendre (10):  A-OBJ, DE-OBJ, OBJ, SUJ| A-OBJ, OBJ, SUJ| A-OBJ, SUJ|
  ATO, DE-OBJ, OBJ, SUJ| ATO, OBJ, SUJ| ATS, P-OBJ, SUJ| ATS, SUJ|
  DE-OBJ, OBJ, SUJ| OBJ, SUJ| P-OBJ, SUJ, refl
```

FIG. 3 – Cadres fonctionnels pour les 4 verbes le plus ambigus : 10 cadres ou plus

direct (plus de la moitié des occurrences), puis ceux à sujet seul (le quart des lemmes), et ceux à objet indirect introduit par *à* ou *de*, et différents types de ditransitifs. Il y a relativement peu de verbes à attribut, mais ils sont très fréquemment utilisés. Dans le tableau 4, les cadres sont représentés avec les fonctions dans l'ordre alphabétique.

L'inconvénient de cette approche est que l'on a perdu des informations par rapport au corpus, en particulier on ne distingue pas les verbes qui prennent seulement un sujet nominal et ceux qui prennent un sujet nominal et phrastique. C'est pourquoi nous passons à une approche mixte.

3.3.2 Extraction de la Sous-catégorisation mixte

Nous avons adopté une représentation mixte, qui garde les fonctions et les catégories. Sans aucune factorisation, ni pour le passif ni pour les clitiques (pronominaux et réfléchis), on obtient 925 cadres différents, avec une moyenne de 3.44 cadres par verbe, et 49% des verbes (668 lemmes) qui n'ont qu'un seul cadre.

Après le dégroupage et l'élimination des duplicats des fonctions décrits dans la sec. 3.2, et avec la normalisation du passif, on réduit le nombre de cadres presque de moitié (on obtient 465 cadres différents), avec une moyenne de 2.78 cadres par verbe. Le nombre de verbes qui ont un seul cadre n'augmente pas beaucoup : 53% de verbes ne sont pas ambigus, soit 727 lemmes.

Nous procédons alors à la factorisation par rapport aux différentes réalisations clitiques (pour les sujets, les objets directs, les *de*-objets et les *à*-objets). Le taux d'ambiguïté baisse à 2.17 et nous obtenons 127 cadres de moins (337), tandis qu'un peu plus de 100 verbes ont un seul cadre

cadre	# types de verbes	occurrences
OBJ, SUJ	986 (72.4%)	6625 (52.9%)
SUJ	346 (25.4%)	1052 (8.4%)
A-OBJ, OBJ, SUJ	184 (13.5%)	423 (3.4%)
A-OBJ, SUJ	136 (9.9%)	423 (3.4%)
DE-OBJ, SUJ	125 (9.1%)	534 (4.3%)
OBJ, P-OBJ, SUJ	101 (7.4%)	165 (1.3%)
DE-OBJ, OBJ, SUJ	98 (7.2%)	196 (1.5%)
P-OBJ, SUJ	81 (5.9%)	215 (1.7%)
ATO, OBJ, SUJ	42 (3.1%)	1929 (15.4%)
SUJ, refl	36 (2.6%)	259 (2.1%)

FIG. 4 – Les 10 cadres de sous-catégorisation fonctionnelle les plus fréquents

	# cadres	moyenne	max. cadres	verbes à un cadre	
				%	#
données brutes	925	3.44	242 (*être*)	49%	668
normalisation passif	465	2.78	114 (*être*)	53.3%	727
normalisation clitiques	337	2.17	76 (*être*)	60.8%	829
normalisation réfléchi	337	2.11	76 (*être*)	62.5%	851
avec OBJ optionnel	338	2.00	75 (*être*)	64.4%	877
sans prépositions	235	1.94	62 (*être*)	64.4%	877

FIG. 5 – Résultats pour les cadres mixtes

`être (62), avoir (23), rester (22), faire (17), passer (14), trouver (14), estimer (13), sembler (13), rendre (13), devenir (11), demander (11), aller (11), porter (11), déclarer (10), laisser (10)`

FIG. 6 – Nombre de cadres mixtes pour 14 verbes les plus ambigus (10 cadres ou plus)

(829). La normalisation par rapport au clitique réfléchi diminue un peu l'ambiguïté (à 2.11 en moyenne) et augmente légèrement le nombre de verbes à un seul cadre (à 851). Le nombre de cadres ne change pas. Nous procédons enfin à la factorisation des objets optionnels, en ajoutant les cadres correspondants, ce qui nous amène à 338 cadres distincts, avec une moyenne de 2 cadres par verbe. Si, enfin, on neutralise la valeur lexicale des prépositions (différentes de *à* ou *de*), on obtient 235 cadres différents au total, et 1.94 en moyenne. Il reste 14 verbes avec plus de 10 cadres, avec un maximum de 62 cadres pour le verbe *être*, indiqués dans le tableau 6. Les résultats sont groupés dans le tableau 5.

Il est clair que l'approche mixte est plus précise mais aboutit à un grand éclatement par rapport à l'approche fonctionnelle : si l'on inclut les catégories, même avec les résultats les plus compactés, nous avons presque 6 fois plus des cadres ! Néanmoins, les taux d'ambiguïtés en moyenne et surtout les nombres de verbes avec un seul cadre sont relativement proches. Ceci nous donne l'espoir que l'approche mixte, plus riche en information, peut être adoptée dans les applications pratiques. Il nous reste certaines factorisations à effectuer : celles qui concerne l'optionalité des autres compléments, et celle qui concerne les attributs. En effet, on distingue les attributs selon leur catégorie, alors que pour *être*, par exemple, il s'agit du même cadre.

cadre	# types de verbes	occurrences
OBJ :NP, SUJ :NP	764 (55.6%)	3489 (27.8%)
SUJ :NP	159 (11.6%)	846 (6.7%)
A-OBJ :PP, OBJ :NP, SUJ :NP	103 (7.5%)	268 (2.1%)
OBJ :Ssub, SUJ :NP	92 (6.7%)	420 (3.3%)
DE-OBJ :PP, SUJ :NP	85 (6.2%)	308 (2.4%)
OBJ :VPinf, SUJ :NP	77 (5.6%)	1636 (13.7%)
P-OBJ :PP, SUJ :NP	73 (5.3%)	170 (1.3%)
OBJ :NP, P-OBJ :PP, SUJ :NP	68 (4.9%)	100 (0.8%)
A-OBJ :PP, SUJ :NP	68 (4.9%)	175 (1.3%)
SUJ :NP, refl :CL	36 (2.6%)	234 (1.9%)

FIG. 7 – Les 10 cadres de sous-catégorisation mixte les plus fréquents

Mais il n'est pas vrai que tous les verbes attributifs acceptent des attributs de n'importe quelle catégorie et il faut sans doute affiner les cadres (Lamiroy & Melis, 2005). On pourrait de même regrouper les compléments phrastiques et infinitifs pour les lemmes qui acceptent les deux.

Si l'on considère les cadres de sous-catégorisation les plus fréquents (tab. 7), on voit que comme dans l'approche précédente, c'est le cadre transitif direct qui arrive en tête. On voit aussi que le complément phrastique concerne plus de lemmes que le complément infinitif, mais beaucoup moins d'occurrences.

3.4 Discussion

Les approches précédentes ne disposant pas de la distinction entre ajouts et arguments dans le corpus de départ, adoptent une approche statistique qui les conduit à ignorer les lemmes à fréquence basse (moins de 5 occurrences). Puisque nous disposons de ces informations dans le corpus, ceci nous permet de considérer aussi les cadres plus rares.

3.4.1 Comparaison avec les travaux sur l'anglais

Pour l'anglais, le taux d'ambiguïté de cadres est rarement mentionné. (Manning, 1993) rapporte la moyenne de 1.43 cadres par verbe pour le lexique de 1856 lemmes, ce qui est comparable avec les chiffres que nous obtenons : un lexique de 1362 lemmes et 1.54 cadres en moyenne (cadres fonctionnels). La différence principale réside non seulement dans la méthode adoptée (l'approche statistique) mais aussi dans le fait qu'il utilise 19 cadres présupposés, tandis que nous les acquérons à partir des annotations dans le corpus. (O'Donovan *et al.*, 2004) adoptent, comme nous, une approche mixte pour la représentation des cadres et obtiennent un lexique de 4362 lemmes, avec environ 4 cadres en moyenne, 38 types de cadres basés uniquement sur les fonctions et 577 cadres acquis si les différents types de prépositions et particules sont inclus. Les auteurs sont quand même obligés à adopter une méthode automatique pour obtenir les annotations fonctionnelles car ces informations ne sont pas présentes dans le corpus.

3.4.2 Comparaison avec les travaux sur le français

Les tables du LADL comportent 38 cadres principaux pour les verbes simples. Ces cadres sont basés sur la catégorie des arguments et non sur les fonctions, et ils tiennent compte de la valeur lexicale de certaines prépositions. Les tables distinguent ainsi parmi les compléments, les cadres où seul un complément infinitif est autorisé (table 1) ou ceux avec un complément nominal introduit par la préposition *à* (table 33). Nos résultats, obtenus par la méthode fonctionnelle, sont presque de même taille (39 cadres différents) mais ils sont en fait différents. Nous avons par exemple des cadres pour les emplois attributifs (cadres avec attribut du sujet ou attribut de l'objet) qui ne figurent pas dans les tables du LADL. D'autre part, nous avons certains emplois figés (cf. verbes avec clitique figé) ce qui crée des cadres supplémentaires. Si l'on compare les tables avec nos résultats obtenus par la méthode mixte, nos chiffres sont très supérieurs. Il y a deux raisons à ceci : d'une part nous avons dégagé des cadres supplémentaires par rapport aux tables LADL (pour les attributs, pour les clitiques figés ou pour les verbes sans sujet) et d'autre part il nous reste encore à faire certains regroupements, par exemple pour les catégories des attributs, ou les emplois à complément infinitif ou phrastique.

(Candito, 1999) et (Abeillé, 2002) décrivent les familles d'arbres de la grammaire FTAG. Il s'agit pourtant de cadres abstraits qui ne sont pas couplés à un lexique de grande taille. Dans la grammaire FTAG sont distinguées 45 familles à tête verbale, dont 15 à arguments nominaux et 24 à arguments phrastiques et 6 à complément adverbial. Cette grammaire inclut, comme ici, un cadre pour les formes verbales sans sujet, *voici* et *voilà*, quelques cadres pour les verbes à clitique figé (comme *s'évanouir* ou *s'appeler* N). Mais il est clair que nous extrayons des cadres supplémentaires.

4 Conclusion

Les résultats préliminaires d'extraction des cadres de sous-categorisation verbale à partir de treebank français présenté dans cet article sont encourageants. Nous avons réussi à considérablement réduire l'ambiguïté de la représentation mixte avec les différentes techniques de factorisations. Ce résultat nous laisse espérer qu'il est donc possible d'incorporer les informations plus riches dans les applications pratiques.

Nous planifions plusieurs extensions de ce travail. Nous allons prendre toutes les occurrences des verbes et tout le corpus annoté en fonctions (15 000 phrases) pour obtenir un lexique plus important. On prévoit aussi l'extraction d'autres catégories prédicatives (cadres de sous-catégorisation des adjectifs ou des noms). Nous pensons incorporer ce lexique dans les applications TAL. Les cadres obtenus peuvent être facilement traduits dans un modèle et un format utilisables par un analyseur syntaxique (par exemple XLFG de L. Clément basé sur LFG), dans une autre application ou selon le format spécifié par le projet LexSynt.[3] Nous proposons aussi de confronter nos résultats à d'autres corpus (par exemple, TLFI ou Frantext) pour valider le lexique et le comparer avec d'autres travaux. Ceci nous permettra aussi enrichir le treebank de départ en ajoutant automatiquement le cadre de sous-catégorisation à chaque occurrence verbale (quand on en a extrait un seul) ou en en ajoutant plusieurs (avec un choix à valider par un annotateur humain) si l'on en a extrait plusieurs.

[3] `http ://lexsynt.inria.fr/index.php`

Références

ABEILLÉ A. (2002). *Une grammaire électronique du français*. CNRS Editions.

ABEILLÉ A. & BARRIER N. (2004). Enriching a French treebank. In *Proceedings of the LREC04 Conference*, Lisbonne.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In *Treebanks*. Kluwer.

BOURIGAULT D. & FRÉROT C. (2005). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*.

BRISCOE T. & CARROLL J. (1993). Generalised probabilistic LR parsing for unifi cation-based grammars. *Computational linguistics*.

CANDITO M.-H. (1999). *Répresentation modulaire et parametrable de grammaires électroniques lexicalisées. Application au français et à l'italien.* PhD thesis, Université Paris7.

CARROLL J. & FANG A. (2004). The automatique acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Conference on Natural Language Processing*, Sanya City, China.

CHESLEY P. & SALMON-ALT S. (2005). Le fi ltrage probabiliste dans l'extraction automatique de cadres de sous-catǵorisation. In *Journé ATALA sur l'interface lexique-grammaire*, Paris.

DANLOS L. (1985). *La generation automatique de textes*. Masson.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2006). Extraction d'information de sous-catégorisation à partir du lexique-grammaire de Maurice Gross. In *TALN 2006*.

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

GUILLET A. & LECLÈRE C. (1992). *La structure des phrases simples en français*. Genève : Droz.

HAN C., YOON J., KIM N. & PALMER M. (2000). *A Feature-Based Lexicalized Tree Adjoining Grammar for Korean*. Rapport interne, IRCS.

HORNBY A. S. (1989). *Oxford Advanced Learner's Dictionary of Current English*. Oxford : Oxford University Press, 4th edition.

LAMIROY B. & MELIS L. (2005). Les copules ressemblent-elles aux auxiliaires ? In SHYLDKROT, H. BAT-ZEEV & N. L. QUERLER, Eds., *Les Périphrases Verbales*, p. 145–170.

MANNING C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31th Meeting of the ACL*, p. 235–242, Columbus, Ohio.

MARINOV S. (2004). Automatic extraction of subcategorization frames for Bulgarian. In P. EGRÉ & L. A. I ALEMANY, Eds., *Proceedings of the Ninth ESSLLI Student Session*.

MEL'CUK I., ARBATCHEWSKY-JUMARIE N. & CLAS A. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques, vol. I, II, III, IV*. Les Presses de l'Université de Montréal.

O'DONOVAN R., BURKE M., CAHILL A., VAN GENABITH J. & WAY A. (2004). Large-scale induction and Evaluation of Lexical Resources from the Penn-II Treebank. In *Proceedings of the 42nd Conference of the Association for Computational Linguistics*, p. 367–374, Barcelona, Spain.

P. PROCTER, Ed. (1978). *Longman Dictionary of Contemporary English*. Burnt Mill, Harlow : Longman.

SAGOT B., CLÉMENT L., DE LA CLERGERIE E. V. & BOULLIER P. (2006). The lefff 2 syntactic lexicon for french : architecture, acquisition, use. In *Actes de LREC 06, Gênes, Italie*.

SARKAR A. & ZEMAN D. (2000). Automatic extraction of subcategorization frames for Czech. In *Proceedings of Colling 2000*.

SURDEANU M., HARABAGIU S., WILLIAMS J. & AARSETH P. (2003). Using predicate-argument structures for information extraction.