

# Le filtrage probabiliste dans l'extraction automatique de cadres de sous-catégorisation

**Paula Chesley**  
Linguistics Department  
University at Buffalo  
Buffalo, NY 14260, USA  
pchesley@buffalo.edu

**Susanne Salmon-Alt**  
ATILF-CNRS  
44, avenue de la Libération  
B.P. 30687  
F-54063 Nancy Cedex  
susanne.alt@loria.fr

## 1 Introduction

Ce travail sur l'extraction automatique des cadres de sous-catégorisation s'inscrit dans le cadre du Lexical Markup Framework (LMF). Comme le verbe constitue une grande partie de la syntaxe d'une phrase, notre travail actuel porte sur la sous-catégorisation de 115 verbes français. Ces verbes ont été choisis pour le travail déjà existant de vérification de leurs sous-catégorisations par la Test Suites for Natural Language Processing (TSNLP). La liste de ces verbes se trouve dans l'annexe A.

Nous avons créé un corpus de 200 occurrences de chacun des 115 verbes dans leurs contextes phrastiques à l'aide de Frantext, une base de textes littéraires disponible en ligne. Ces 200 occurrences ont été choisies de façon aléatoire. Des travaux apparentés optent pour des nombres d'occurrences de chaque verbe bien inférieurs au nôtre : Brent (1993) remarque qu'il n'y a pas de différence significative des résultats lorsque le nombre d'occurrences varie entre 50 et 150 (p. 256), et Manning et Schütze (1999) parlent de résultats fiables lorsqu'il y a 80 occurrences d'un verbe donné (p. 275). Nous estimons donc que 200 occurrences de chaque verbe peut donner des résultats représentatifs de la sous-catégorisation des verbes dans notre corpus.

Ensuite, nous avons effectué une analyse syntaxique profonde (en dépendances et constituants) de ces phrases<sup>1</sup>. À partir de ces analyses, nous avons cherché les cadres de sous-catégorisation des 115 verbes afin de les mettre dans le format des données du LMF. À l'heure actuelle, nous examinons des méthodes probabilistes de filtrage des cadres. L'hypothèse qui sous-tend ce filtrage suppose qu'à partir d'une analyse syntaxique qui comprend déjà un lexique, on puisse extraire davantage d'informations sur le lexique, et que ce dernier sera amélioré par le filtrage. Nos premiers résultats sont encourageants.

<sup>1</sup>Des informations sur l'analyseur, celui de VISL, se trouve en ligne à l'URL suivant :  
[http://visl.hum.sdu.dk/visl/fr/info/taginfo\\_french.html](http://visl.hum.sdu.dk/visl/fr/info/taginfo_french.html).

## 2 Extraction des cadres de sous-catégorisation

La terminologie évoquée dans la plupart de travaux sur la sous-catégorisation mérite d'être discutée. Le sens du terme *cadre de sous-catégorisation* peut varier selon l'auteur, à savoir, les éléments sous-catégorisés, comme des prépositions, constituent-ils un nouveau cadre (e.g., "arriver [<sub>pp</sub>prep à]"), ou une instanciation d'un cadre "squelette" (e.g., "arriver [<sub>pp</sub>]")? Dans ce travail nous adopterons la première approche, comme le fait Manning (1993), tout en sachant que c'est un point discutable. De plus, on mentionne souvent la distinction argument-circonstant<sup>2</sup>. À cet égard, nous soulignons que le travail actuel ne porte que sur la syntaxe et la sous-catégorisation des verbes. Nous ne distinguons les cadres de sous-catégorisation des compléments circonstanciels que par la fréquence de co-occurrence d'un verbe avec un cadre, ces derniers ayant, bien sûr, une fréquence de co-occurrence plus élevée que le verbe et un complément circonstanciel.

L'analyseur syntaxique VISL comprend un lexique. Or, les cadres de sous-catégorisation faisant partie des entrées verbales dans ce lexique, il semblerait que notre travail ne consiste qu'en changeant le format des cadres de sous-catégorisation de la sortie de l'analyseur à celui du LMF. Toutefois, la sortie de l'analyseur témoigne des difficultés à bien analyser le rattachement prépositionnel, les compléments subordonnés, et les compléments d'objet indirects de certains verbes.

Le lexique et le jeu d'étiquettes de VISL permettent de savoir si les constituants mentionnés ci-dessus ont été considérés, en théorie, comme étant des éléments sous-catégorisés ou non. Par exemple, l'analyse d'un groupe prépositionnel dont la préposition est sous-catégorisé par le verbe devrait être soit un objet datif, également appelé un complément d'objet indirect (selon les étiquettes de VISL, un "Odat"/ "Oi"), soit un objet prépositionnel

<sup>2</sup>Nous remercions un lecteur anonyme de nous avoir rappelés ce point.

(“Op”). Or, l’analyse de VISL dans ces cas est souvent celle d’un adjectif adverbial (“fAdv”, pour *free adverbial*). Le lexique encore incomplet de l’analyseur et la difficulté d’effectuer une analyse juste pour des phrases longues, c’est-à-dire, plus de 20 mots, sont les raisons principales pour ces mauvaises analyses. Ces difficultés ont motivé nos expériences sur le filtrage afin de réduire le nombre de cadres erronés.

### 3 Le filtrage probabiliste des cadres

Si les erreurs de l’analyseur ne sont pas toujours bonnes, elles sont les meilleures erreurs auxquelles on peut s’attendre : il est plus facile de chercher toutes les occurrences des “fAdv”, qui sont toujours au même niveau dans la structure arborescente de l’analyse, que d’aller à la chasse de tous les groupes prépositionnels rattachés à un substantif quelconque dans l’arborescence de la phrase. Si la mauvaise analyse des cadres prépositionnels en tant que “fAdv” est systématique (ce que nous avons observé informellement), le filtrage probabiliste nous indiquerait quand même si une préposition constitue un cadre pour un verbe donné. Ainsi avons-nous choisi de retenir ces mauvaises analyses de VISL dans le format du LMF, pour les filtrer dans une étape suivante.

La méthode de filtrage dont nous nous servons, celle de la distribution binomiale des cadres, est aussi celle de Manning (1993), de Brent (1993), et de Briscoe et Carroll (1997), entre autres. Nous appelons un *indice*, tout cadre que nous avons au départ, sans savoir s’il est bien analysé. La distribution binomiale examine donc la différence entre le nombre de co-occurrences d’un indice et d’un verbe et le nombre de fois que l’on voit ce dernier dans le corpus entier. Plus cette différence est importante, moins il est probable que le verbe sous-catégorise l’indice. Soient  $m$ , le nombre total d’occurrences de verbes dans le corpus,  $n$ , le nombre de fois le verbe apparaît avec l’indice, et  $L_c$ , la limite maximale estimée de la probabilité qu’un verbe qui ne sous-catégorise pas un cadre  $c$ , apparaisse pourtant avec l’indice de  $c$  (Manning (1993), p. 4). Si on adopte un seuil de .01 au-dessus duquel l’indice n’est pas sous-catégorisé par le verbe<sup>3</sup>, la distribution binomiale détermine, pour chaque co-occurrence d’un verbe et d’un indice, la probabilité que l’indice soit faux, c’est-à-dire que l’indice ne constitue pas un cadre de sous-catégorisation du verbe :

$$(1) \quad \sum_{i=n}^m \frac{m!}{i!(m-i)!} L_c^i (1-L_c)^{m-i}$$

Manning fixe la limite  $L_c$  de façon manuelle pour chaque cadre, alors que Brent (1992) se sert d’une façon automa-

<sup>3</sup>Pour Manning, ce seuil est fixé à .02. Pour Briscoe et Carroll (1997), le chiffre équivalent est .05.

tique pour déterminer  $L_c$ . C’est cette dernière approche que nous avons adoptée dans notre travail.

Pour résumer, l’algorithme de Brent consiste en examiner toute occurrence d’un cadre  $c$  avec tout verbe au delà d’une certaine fréquence, qui est de 30 occurrences dans ce travail. A partir de ces occurrences on construit un histogramme fondé sur le nombre de co-occurrences des indices avec les verbes de fréquence importante, dans lequel on cherche une distribution binomiale à l’extrémité inférieure qui signifie les mauvais indices de  $c$ . La moyenne dans cette distribution est une bonne estimation du taux  $L_c$  de mauvais indices.

Afin de découvrir cette distribution binomiale, on doit tout d’abord chercher un intervalle  $j_0$  de l’histogramme qui est la limite maximale de la distribution. Ce faire revient à faire l’hypothèse que chaque intervalle  $j$  dans l’histogramme est  $j_0$ . L’intervalle est alors évalué en comparant les distributions attendue et observée du cadre à chacun des  $j$  intervalles dans l’histogramme. Le meilleur intervalle  $j_0$  est celui pour lequel il y a la moindre différence entre ces deux distributions. En pratique, puisque le nombre d’occurrences pour chaque verbe est différent, on doit établir, pour chaque intervalle dans l’histogramme, un poids proportionnel de chaque verbe.

### 4 Evaluation

Dans un premier temps, nous nous concentrons sur le filtrage des cadres prépositionnels et les cadres de compléments d’objet directs. Voici, brièvement, nos résultats de certaines combinaisons de verbes avec ces cadres par rapport aux résultats initiaux de l’analyseur de VISL :

1. **arriver**. Nos résultats soutiennent l’hypothèse que ce verbe sous-catégorise les prépositions “à” et “de”, avec des taux de mauvais indices de  $1.18958149731991e-41$  et de  $0.00456307675498892$ , respectivement. Notons que ces résultats supposent que la première préposition est plus susceptible d’être un vrai cadre que la dernière.
2. **diriger**. La sortie de VISL n’indique pas que ce verbe sous-catégorise la préposition “vers”. Cette préposition ayant un taux de mauvais indices de  $5.23779295890851e-53$ , nous supposons que “vers” constitue un cadre du verbe.
3. **donner**. Nos résultats présument que la préposition “de”, avec un taux de mauvais indices de  $5.99543182649634e-22$ , serait un cadre du verbe; la sortie de VISL indique le contraire. Nous pensons que c’est un point discutable.
4. **regretter**. Pour ce verbe nos résultats ne diffèrent de façon signative de ceux de l’analyseur de VISL.

Souvenons que ce premier travail ne prend pas en compte les compléments subordonnés, que “regretter” sous-catégorise. Nous pensons que notre programme montrerait ce cadre de façon définitive.

Nous avons vérifié de façon manuelle les résultats de ce programme sur 10 verbes choisis pour leurs cadres de sous-catégorisation hétérogènes ainsi que pour leurs fréquences importantes. Ces verbes se trouvent dans l’annexe B. Pour ces verbes, nous avons relevé 11 nouveaux types de prépositions ou de compléments d’objet directs que l’analyseur au départ n’avait pas pris en compte. Certains, comme la préposition “vers” du verbe “diriger”, doivent clairement être des éléments sous-catégorisés, alors que d’autres, comme “de” pour “donner”, sont discutables. Peut-être ce résultat vient-il d’une mauvaise analyse syntaxique en tant que le partitif “de”, ou d’un seuil  $L_c$  à mieux cerner.

Les travaux d’évaluation de Brent (1993) et de Manning (1993) pour l’anglais témoignent de l’efficacité de la combinaison d’un premier traitement, comme celui de notre analyseur, suivi du filtrage par la distribution binomiale des indices apparaissant avec un verbe donné. Ces deux auteurs examinent des *types* plutôt que des *tokens* pour faire l’évaluation<sup>4</sup>, et le premier voit des taux de précision de 96% à 100%, selon le cadre (p. 255), alors que le deuxième note une limite minimale de précision de 90% (p. 6). Il faut noter que Brent se borne à la recherche de six cadres assez fréquents ; en revanche, Manning s’intéresse à l’extraction de 19 cadres. Les taux de rappel de Brent sont de 47% à 100% selon le cadre. Manning estime une limite minimale du rappel de 82%, cette fois-ci sur les *tokens*. Ne souhaitant pas nous borner à une vingtaine de cadres, nous avons pensé à la possibilité que nos résultats soient légèrement moins robustes que ceux de ces deux auteurs.

## 5 Conclusion et perspectives

L’objectif principal de cette expérience était de mettre au point et de tester une méthode probabiliste pour l’acquisition de cadres de sous-catégorisation à partir des résultats d’une analyse syntaxique automatique profonde en constituants et dépendances. C’est surtout ce dernier point qui la distingue des expériences précédentes : celles-ci ont été menées sur des corpus bruts ou étiquetés seulement Manning (1993), en utilisant des analyseurs de surface Briscoe et Carroll (1997), ou avec un analyseur

<sup>4</sup>Faire une évaluation des *types* plutôt que des *tokens* ne suppose aucune répétition de verbes ou de combinaisons verbe-cadre. On évite donc des taux élevés à cause de la répétition des plus communs verbes avec les plus communs cadres. Manning et Schütze (1999) remarque que Brent se sert des *tokens* pour son évaluation (p. 274). Peut-être cette différence vient-elle de la façon dont Brent explique sa démarche d’évaluation.

en dépendances Frérot et al. (2003). Dans notre configuration, la question était de savoir si des analyses probabilistes sur des résultats d’analyses complètes (mais potentiellement fausses) pourraient contribuer à améliorer le lexique initial de l’analyseur. La réponse à cette question est positive : pour les 10 verbes les plus fréquents de notre jeu de test, nous avons relevé 11 nouveaux types de constituants potentiellement sous-catégorisables, et ce en travaillant uniquement sur les groupes prépositionnels et les compléments d’objet direct, en adoptant une valeur de seuil relativement sévère par rapport aux travaux précédents. Une des questions ouvertes reste la question de l’évaluation des résultats : en l’absence de lexiques de référence, on peut s’en tenir à des dictionnaires classiques ou à un ensemble de tests linguistiques (suppression, déplacement, substitution), mais une part de subjectivité dans le jugement ne peut pas être complètement écartée.

Dans le même esprit de recherche, nous devons avant tout faire des expériences avec diverses valeurs de la limite  $L_c$ , la limite maximale estimée de la probabilité qu’un verbe qui ne sous-catégorise pas un cadre, discutée dans la section 3. Nous avons choisi une valeur de .01 en regardant la sortie de certains exemples, une valeur relativement basse par rapport à celles d’autres chercheurs évoquées plus haut. Brent (1993) remarque que plus ce seuil est bas, plus on peut se faire confiance dans la conclusion de la présence d’un cadre de sous-catégorisation (p. 251). Cependant, il constate que les meilleurs résultats apparaissent lorsque ce seuil varie entre .01 et .05 selon le cadre (p. 256). Aussi tenons-nous à faire les mêmes expériences pour tous les cadres éventuels, par exemple le complément subordonné “que”.

Malgré une performance efficace, les méthodes d’extraction de l’information lexicale à partir de corpus ont des limites. Nous soulignons tout d’abord qu’à ce jour notre expérience n’est pas capable de doter le lexique de VISL d’informations sémantiques sur les arguments des verbes, telles que *patient*, *agent*, etc. Des travaux d’extraction à partir des ressources telles que le TLFi ou les tables du LADL, peuvent prendre en compte ces informations. Pour pouvoir extraire des informations sémantiques à partir de corpus, il faudrait un travail sur l’interface syntaxe-sémantique tel qu’une étude sur les classes de verbes en français. Nous nous intéressons à ce travail car l’information que l’on retrouve dans les ressources lexicales mentionnées plus haut peut être incomplète, et parfois peu détaillée.

La plupart des travaux cités ont un taux de rappel inférieur au taux de précision, ce qui indique qu’il y a souvent un manque de données pour les verbes et les cadres de basse fréquence. Nous reconnaissons aussi qu’actuellement notre expérience n’est capable de prendre en compte ni les homonymes ni les polysèmes, ce qui est

le cas pour plusieurs travaux de corpus. En revanche nous pensons que les divers cadres de sous-catégorisation pourraient aider à distinguer divers sens d'un même graphique verbal. Prenons par exemple le verbe "assurer" : on dit "assurer quelqu'un de quelque chose" et aussi "assurer en une matière". Ces polysèmes ont différents cadres de sous-catégorisation, et nous pourrions exploiter cette information dans de futures recherches sur les homonymes et les polysèmes.

Nous récapitulons l'observation de Frérot et al. (2003), un travail qui porte sur un sujet apparenté au nôtre : "la plupart des travaux cités portent sur la langue anglaise" (p. 16). En entreprenant le rattachement prépositionnel, Frérot et al. ont non seulement enrichi des ressources lexicales pour le français, mais ont également étudié un phénomène récurrent dans plusieurs langues. Comme nous avons d'une part automatisé le seuil  $L_c$  de mauvais indices et d'autre part utilisé un analyseur en constituants et dépendances, nous espérons que le travail actuel complète le leur, et, de façon générale, qu'il enrichira les études sur la sous-catégorisation dans les ressources lexicales, y compris le LMF et l'analyseur de VISL que nous avons utilisés au départ, pour le français.

## Bibliographie

- M. Brent. 1992. Robust Acquisition of Subcategorization Frames from Unrestricted Text : Unsupervised Learning with Syntactic Knowledge. Master's thesis, Johns Hopkins University, Baltimore, MD.
- M. Brent. 1993. From grammar to lexicon : Unsupervised learning of lexical syntax. *Computational Linguistics*, 19 :243–262.
- T. Briscoe et J. Carroll. 1997. Automatic Extraction of a Subcategorization from Corpora. pages 356–363. Proceedings of the 5th ACL Conference on Applied Natural Language Processing.
- C. Frérot, D. Bourigault, et C. Fabre, 2003. *Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition « de »*, volume 44 :3, pages 1–20.
- C. Manning et H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.
- C. Manning. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. pages 235–242. Proceedings of the 31st ACL.

## A Les 115 verbes du corpus

aborder	croire	lire	regretter
accepter	croître	livrer	représenter
acheter	decider	maintenir	requérir
agir	démarrer	manger	réserver
aider	devenir	marcher	restaurer
aimer	devoir	marier	rester

aller	dire	mentir	rêver
apercevoir	diriger	mettre	savoir
apparaître	diviser	montrer	séparer
appeler	donner	offrir	signer
apprendre	dormir	ouvrir	sommer
arriver	durer	ouvrir	sortir
asseoir	écrire	paraître	sucrer
avertir	entendre	parler	suer
avoir	entreprendre	participer	suffire
avouer	entrer	partir	suivre
boire	espérer	passer	supposer
causer	étayer	penser	taire
cesser	être	permettre	terminer
combattre	exceller	persuader	tomber
commencer	faillir	plaire	toucher
comparer	faire	pleuvoir	transférer
comprendre	falloir	prendre	travailler
connaître	fontionner	présenter	trouver
constituer	hésiter	prononcer	venir
contrer	indiquer	proposer	vivre
convaincre	intéresser	provoquer	voir
courir	interroger	raconter	vouloir
craindre	laisser	recevoir	

## B Les 10 verbes du jeu de test

arriver  
courir  
diriger  
donner  
faire  
falloir  
manger  
plaie  
pleuvoir  
regretter