SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

La structure argumentale des noms déverbaux : du corpus au lexique et du lexique au corpus

Condette, Marie-Hélène, Marín, Rafael, & Merlo, Aurélie

Université Lille 3 & CNRS (UMR 8163)
marie-helene.condette@wanadoo.fr, rafael.marin@univ-lille3.fr, aurelie.merlo@yahoo.fr

1 Introduction

Dans ce travail, nous présentons une ressource linguistique incluant un lexique –extrait d'un corpus– de noms déverbaux codés par rapport à sa structure argumentale (SA), qui est, à notre connaissance, la seule ressource de ce type pour le français.

Actuellement, il existe pour le français des ressources contenant des informations sur la SA de noms prédicatifs telles que le *Lexique-Grammaire* des noms prédicatifs (Gross, 1975) et le *Lexique Actif du Français* (LAF) (Mel'čuk & Polguère, 2007). Il existe également des ressources comportant des informations de SA verbale ; entre autres : *Lexique-Grammaire* (Leclère, 2002) ; *LexValf* (Salkoff & Valli, 2005) ; *SynLex* (Gardent & *al.*, 2006) ; *DicoValence* (van den Eynde & Mertens, 2006) ; *Lefff* (Sagot & *al.* 2006) ; *TreeLex* (Kupść & Abeillé, 2008) ; *EasyLex* (Gardent, 2009) ; *LexSchem* (Messiant, Korhonen & Poibeau, 2008) ; *LGLex* (Tolone, 2011). Néanmoins, il n'existe pas de ressource alliant à la fois SA nominale et verbale comme le propose le lexique de noms déverbaux que nous allons présenter.

Pour d'autres langues, il existe des ressources sur la SA des noms déverbaux comme *Nomlex* (Macleod & *al.*, 1998) et *NomBank* (Meyers & *al.*, 2004) pour l'anglais ou *Ancora* (Taulé & *al.*, 2008) pour l'espagnol.

Dans notre cas, nous ne nous intéressons pas à un système de très large couverture, mais à un prototype, permettant de confronter certains postulats théoriques à un échantillon significatif des données réelles. Quelque part, on pourrait dire que notre approche est plus qualitative que quantitative. En cela, le lexique que nous présentons n'est donc pas une ressource conçue pour le Traitement Automatique des Langues (TAL).

Parmi les questions théoriques que l'on veut vérifier, il y a l'hypothèse de la préservation de la SA (HPSA), selon laquelle les noms déverbaux héritent leur SA des verbes dont ils dérivent.

2 Ressources existantes

2.1 Ressources contenant des informations sur la structure argumentale verbale

On peut relever un grand nombre de ressources en français comprenant des informations sur la SA verbale. Certaines de ces ressources ont été créées soit manuellement soit semi-automatiquement à partir de lexiques telles que le *Lexique-Grammaire* (Gross, 1975 ; Leclère, 2002), *LexValf* (Salkoff & Valli, 2005), *DicoValence* (van den Eynde & Mertens, 2006), *Lefff* (Sagot & *al.,* 2006), *SynLex* (Gardent & *al.*, 2006) et *LGLex* (Tolone, 2011). D'autres ressources ont été obtenues automatiquement à partir de corpus telles que *LexSchem* (Messiant, Korhonen & Poibeau, 2008), *TreeLex* (Kupść & Abeillé, 2008) et *EasyLex* (Gardent, 2009). Voici la description de quelques-unes de ces ressources comprenant des informations sur la structure argumentale verbale. Cette description nous permettra par la suite de montrer l'apport de notre ressource à ce sujet.

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20120100271

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

Le *Lexique-Grammaire* (Gross, 1975 ; Leclère, 2002) contient 67 tables regroupant 5738 verbes[1]. Chaque table regroupe des phrases simples ayant une construction syntaxique commune et contient des informations morpho-syntaxiques et sémantiques sur les arguments. Le codage des informations contenues dans les tables du *Lexique-Grammaire* ne permet pas d'utiliser cette ressource en Traitement Automatique des Langues (TAL). Les ressources qui ont suivi, réalisées à partir du *Lexique-Grammaire*, comme *DicoValence* (van den Eynde & Mertens 2006), *SynLex* (Gardent & *al.,* 2006), *Lefff* (Sagot & *al.,* 2006) ou *LGLex* (Tolone 2011), avaient pour objectif de proposer des informations sur la structure argumentale verbale exploitables en TAL.

Dans le cadre de sa thèse, Messiant (2010) s'est penché sur l'acquisition automatique en corpus[2] brut d'informations lexicales. Il s'est particulièrement intéressé à l'élaboration d'un système d'acquisition automatique de schémas de sous-catégorisation (SSC) verbale. Un schéma de sous-catégorisation est défini par Messiant (2010 : 12) comme étant un « phénomène syntaxique qui dénote la tendance des prédicats à imposer à leur entourage des configurations syntaxiques particulières. Ces configurations sont représentées par des schémas (ou cadres) de sous-catégorisation ». La sous-catégorisation verbale est décrite en termes syntaxiques[3] :

(1) $(Julie)_{SUJ/SN}$ *a donné* $(un\ livre)_{OBJ/SN}$ $(à\ Marc)_{P\text{-}OBJ/SP}$

[SUJ : SN, OBJ : SN, P-OBJ : SP]

Le système d'acquisition automatique de SSC de Messiant (2010) a permis l'élaboration automatique du lexique *LexSchem* (Messiant, Korhonen & Poibeau, 2008), un lexique de sous-catégorisation verbale à large couverture pour le français. Ce lexique contient actuellement 9476 couples verbes-SSC, 4656 lemmes verbaux et 107 SSC différents[4].

Les lexiques *TreeLex* (Kupść & Abeillé 2008) et *EasyLex* (Gardent 2009) contiennent également des informations sur la sous-catégorisation verbale. *TreeLex* est un lexique libre[5] de sous-catégorisation verbale extrait automatiquement du corpus arboré de Paris 7 (Abeillé & *al.,* 2003) comportant environ 2000 lemmes verbaux et 180 SSC. *EasyLex* est né du projet TALC (Traitement Automatique des Langues et des Connaissances) et des travaux de Claire Gardent. Ce lexique comporte 4800 verbes et propose en moyenne 6 SSC par verbe.

Voici un tableau récapitulatif des ressources citées comportant des informations sur la structure argumentale verbale.

Ressource	Nombre de lemmes	Exploitable en TAL
Lefff	6825	oui
Lexique-Grammaire	5738	non
LGLex	5694	oui
SynLex	5244	oui
EasyLex	4800	oui
LexSchem	4656	oui
DicoValence	3700	non
TreeLex	2000	oui
LexValf	975	oui

Tableau 1 : récapitulatif des ressources comportant des informations sur la SA verbale

2.2 Ressources contenant des informations sur la structure argumentale de noms déverbaux

Concernant le français, il existe actuellement deux ressources contenant des informations sur la SA nominale mais ces ressources ne sont pas destinées spécifiquement aux noms déverbaux. Ces ressources sont le *Lexique-Grammaire* des noms prédicatifs (Gross, 1975) et le *Lexique Actif du Français* (Mel'čuk & Polguère, 2007). Le *Lexique-Grammaire* des noms prédicatifs (Gross, 1975) est constitué de 81 tables contenant des informations morpho-syntaxiques et sémantiques sur les arguments. Ainsi, pour le nom *admiration*, nous trouvons cette description prédicative : *le N de N0 Prép N1* où *N1* peut-être un *Nhum* ou un *N-hum*. Le *Lexique Actif du Français* (Mel'čuk & Polguère, 2007) est construit selon la théorie de la Lexicologie Explicative et Combinatoire (Mel'čuk & *al.*, 1995). Cette théorie établit un nombre restreint de patrons de liens lexicaux dérivationnels et collocationnels appelés fonctions lexicales dans la terminologie Sens-Texte (instrument, intensificateur, etc.) : c'est précisément la modélisation des liens lexicaux établis en termes de fonctions lexicales propre à cette théorie qui est présentée dans les articles du *LAF*. Cette modélisation implique notamment la formalisation des actants sémantiques sous la forme X, Y, Z, qui est expliquée dans le chapitre 1 du *LAF* (Mel'čuk & Polguère, 2007 : 24) :

> « Le sens de la grande majorité des lexies ne peut être clairement compris, et donc décrit, sans prendre en compte les « participants » des situations que ces lexies désignent, qui s'expriment dans la phrase sous le contrôle des lexies en question. Nous appellerons ces participants *actants sémantiques* et nous les identifierons dans les articles du LAF par les variables X, Y… »

Par exemple, dans le *LAF*, pour le nom masculin ABAISSEMENT (*ibid*., : 83-84), on trouve ainsi :

- pour le sens I.1 DIMINUTION, la modélisation « abaissement du paramètre X [= *de* N, A_{poss}] » ;
- pour le sens I.2 FAIT DE DÉPLACER, la modélisation « abaissement par l'individu X [= *par* N] de l'entité Y [= *de* N, A_{poss}] » ;
- pour le sens II FORME, la modélisation « abaissement de l'entité X [= *de* N, A_{poss}] ».

De même, pour l'entrée 1 ou l'unité lexicale 1 du nom masculin ABATTEMENT[1], au sens de ÉTAT D'ESPRIT, on a la modélisation « abattement de l'individu X [= *de* N, A_{poss}] » (*ibid*., : 87), et pour l'unité lexicale 2 de ce même nom ABATTEMENT[2], au sens de FAIT DE FAIRE DIMINUER, on trouve la formalisation « abattement par la personne X [= *de* N, *par* N, A_{poss}] de la somme d'argent Y [= *de* N, *de* Num N] à la personne Z [= *à* N, *pour* N, A_{poss}] sur W [= *sur* N] » (*ibid*., : 87).

Concernant les autres langues, *Nomlex* (Macleod & *al.*, 1998) et *Nombank* (Meyers & *al.*, 2004) sont deux projets de production d'un lexique de noms prédicatifs en anglais. Le projet *AnCora* (Taulé & *al.*, 2008) est un projet de lexique en langue espagnole et catalane comprenant 500 000 mots annotés à différents niveaux : morphologique (partie du discours), syntaxique (constituants et fonctions) et sémantique (structure argumentale, rôle thématique, classe sémantique, entités nommées et sens WordNet).

Nomlex[6] (NOMinalization LEXicon) est un lexique de nominalisations anglaises développé dans le cadre du Proteus Project par l'Université de New York sous la direction de Catherine MacLeod. Le but de ce projet vise à déterminer quels sont les compléments autorisés pour une nominalisation et à mettre en relation les compléments nominaux et les arguments du verbe correspondant, autrement dit, à établir un lien entre les arguments d'une nominalisation et la structure argumentale prédicative du verbe de base. Sur le plan du contenu, le projet inclut, d'une part, la prise en compte des principaux arguments du verbe (sujet, complément direct, complément indirect) ainsi que certains compléments verbaux plus secondaires directement liés aux compléments nominaux et, d'autre part, l'élaboration d'une entrée de nominalisation étendue, incluant des informations relatives aux verbes support qui accompagnent souvent les nominalisations (ex. *lancer une attaque*, *faire une promenade*). Le projet *Nomlex* comprend 1 025 entrées lexicales des nominalisations les plus fréquentes issues de différents corpus (entres autres, Brown Corpus, Wall Street Journal). Dans le cadre du projet *Nomlex*, il est également prévu d'annoter toutes les

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

nominalisations issues d'un autre corpus, le *Penn Treebank*, afin d'étendre et de valider les entrées de *Nomlex*.

NomBank[7] est un projet d'annotation sur corpus de l'Université de New York, en lien avec le projet *PropBank*[8] de l'Université de Colorado. L'objectif de *NomBank* est d'analyser les arguments des noms dans le PropBank Corpus, qui est constitué par le Wall Street Journal Corpus du *Penn Treebank*, tout comme *PropBank* vise à y étudier les arguments des verbes. Dans le cadre du processus d'annotation, le projet *Nombank* produit un certain nombre de ressources, dont divers dictionnaires, permettant d'étiqueter les divers arguments et les adjoints des noms candidats, avec l'attribution de rôles en accord avec les parties du discours. Ce projet a commencé en liaison avec le projet *Nomlex* de Catherine MacLeod. Dans cette optique, l'objectif de *Nombank* est de définir et de décrire la structure argumentale des noms de la manière la plus fine et la plus détaillée possible, ce qui implique l'analyse de divers phénomènes tels que les constructions des verbes support, les arguments des copules, les constructions des syntagmes prépositionnels : un des intérêts de cette étude est de constater que l'argument d'un nom peut se trouver en dehors du syntagme nominal dont ce nom est la tête. Ce projet vise ainsi à analyser les nominalisations des verbes mais aussi celles des adjectifs. La version 1.0 de *NomBank* est sortie le 17 décembre 2007 : elle couvre tous les noms analysables du Wall Street Journal Corpus du *Penn Treebank*, à savoir, 114 576 propositions et 202 965 occurrences de noms.

Le lexique obtenu dans le cadre du projet *AnCora* (Taulé & *al.* , 2008) a été élaboré à partir d'annotations manuelles et semi-automatiques effectuées à différents niveaux. Au niveau sémantique, l'annotation de la structure argumentale verbale a permis l'enrichissement du niveau syntaxique, comme l'atteste le tableau ci-dessous récapitulant les fonctions possibles que peut réaliser chaque argument.

Arg0	Agent complement, Direct object, Indirect object, Subject
Arg1	Adjunct, Direct object, Prep. comp., Subject
Arg2	Attribute, Adjunct, Direct object, Indirect object, Predicative, Prep. comp., Subject
Arg3	Adjunct, Indirect object, Predicative
Arg4	Adjunct
ArgM	Adjunct, Predicative
ArgA	Prep. comp., Subject
ArgL	Adjunct, Direct object, Predicative, Prep. comp., Subject

Tableau 2 : récapitulatif des fonctions possibles que peut réaliser chaque argument dans la ressource AnCora

L'annotation de la structure thématique dans le projet AnCora a révélé que chacun de ces arguments pouvait coïncider avec des rôles thématiques spécifiques comme les suivants : AGT (Agent), AGI (Induced Agent), CAU (Cause), EXP (Experiencer), SCR (Source), PAT (Patient), TEM (Theme), ATR (Attribute), BEN (Beneficiary), EXT (Extension), INS (Instrument), LOC (Locative), TMP (Time), MNR (Manner), ORI (Origin), DES (Goal), FIN (Purpose), EIN (Initial State), EFI (Final State) et ADV (Adverbial). Enfin, concernant l'annotation au niveau sémantique dans le projet AnCora, une annotation manuelle a consisté en l'attribution d'un sens à chaque substantif à partir de WordNet.

La ressource que nous allons présenter ici s'inspire particulièrement des ressources *Nomlex, NomBank* et *AnCora* pour proposer un lexique français des noms déverbaux contenant des informations sur leur structure argumentale.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

3 Notre ressource

La ressource que nous présentons ici, issue du projet Nomage (Balvet *et al.*, 2011) vise la description de la SA d'un lexique de noms déverbaux provenant d'un corpus électronique annoté, le *French Treebank* (Abeillé, 2003).

3.1 Méthodologie

La méthodologie générale appliquée pour le développement de cette ressource a consisté globalement à effectuer les étapes suivantes :

- Extraire du corpus les phrases contenant des noms déverbaux en s'appuyant notamment sur les suffixes identifiés comme tels (*-age*, *-ment*, *-tion*, etc.) ;
- Élaborer un lexique de ces noms (656 lemmes et 742 acceptions) à partir des occurrences des noms déverbaux relevés (4 018) ;
- Élaborer un lexique des verbes de base (648 lemmes et 677 acceptions) à partir du lexique des noms déverbaux ;
- Associer une SA à chaque acception des noms déverbaux répertoriés et à chaque acception des verbes de base ;
- Vérifier la réalisation effective de la SA des noms dans ses occurrences respectives en corpus.

Ainsi, un des apports spécifiques de notre ressource est de permettre la comparaison entre la SA des noms déverbaux et celle des verbes de base.

Notre approche n'est donc pas « corpus-based » : nous n'élaborons pas la SA des nominalisations à partir de ce que nous trouvons concrètement dans le corpus (comme le font par exemple Kupsc (2009) ou Messiant (2010)). À l'inverse, nous déterminons d'abord la SA théorique ou « idéale » des nominalisations pour ensuite étudier et examiner quelle est sa réalisation effective « de surface » en corpus, afin de pouvoir notamment mettre en évidence, comparer, voire mesurer, le degré d'adéquation, de correspondance ou au contraire le décalage éventuel existant, d'un côté, entre la SA des noms déverbaux et celle des verbes de base et, d'un autre, entre lexique et corpus ; autrement dit entre théorie et données.

3.2 Questions théoriques

Nous partons de l'hypothèse de la préservation de la SA (HPSA), selon laquelle la SA des noms déverbaux est héritée ou dérivée de celle des verbes de base. Il ne faut pas interpréter la HPSA de façon stricte en postulant que la totalité de la SA des noms déverbaux est héritée de leur base verbale, puisqu'il peut y avoir certains décalages lors du passage du verbe au nom : pertes d'arguments, changement de prépositions, etc. En revanche, la SA du nom, en tant que SA dérivée, ne peut pas contenir d'arguments qui n'étaient pas déjà inclus dans la SA du verbe : « Nullum argumentum est in nomine quod non prius erat in verbo ».

Avec la recherche que nous présentons ici, nous voulons essayer de vérifier d'autres postulats, comme par exemple celui de Grimshaw (1990) selon lequel l'argument qui correspond à l'Objet Direct (OD) du verbe est obligatoire chez les nominalisations.

En outre, comme cela a déjà été suggéré, nous voulons aussi vérifier si les arguments des nominalisations sont beaucoup plus souvent optionnels que ceux des verbes (Peris & Taulé, à par.).

4 L'annotation de la SA

En ce qui concerne la codification de la SA, nous avons commencé par le lexique des verbes de base, en prenant comme point de départ les paradigmes de *DicoValence* (van den Eynde & Mertens, 2003). Pour

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

décrire la SA des noms, nous nous sommes également inspirés de la Lexicologie Explicative et Combinatoire (LEC) de la Théorie Sens-Texte (Mel'čuk & Polguère, 2007).

4.1 DicoValence (van den Eynde & Mertens, 2006)

Sur le plan de la méthodologie, pour décrire la structure argumentale du verbe de base dont dérive la nominalisation, nous sommes partis de *DicoValence* (van den Eynde & Mertens, 2003)[9], un lexique décrivant la SA de plus de 3 700 verbes simples du français[10], élaboré dans le cadre de l'approche pronominale (Blanche-Benveniste *et al*., 1984).

Dans *DicoValence*, la description de la SA est assez fine : aux 3 700 verbes correspondent plus de 8 000 cadres valenciels. Elle se fait de la façon suivante :

> « [D]'abord, pour chaque place de valence (appelée « paradigme ») le dictionnaire précise le paradigme de pronoms qui y est associé et qui couvre « en intention » les lexicalisations possibles ; ensuite, la délimitation d'un cadre de valence, appelé « formulation », repose non seulement sur la configuration (nombre, nature, caractère facultatif, composition) de ces paradigmes pronominaux, mais aussi sur les autres propriétés de construction associées à cette configuration, comme les « reformulations » passives. »[11].

Les paradigmes les plus courants et les plus utilisés dans un schéma valenciel sont[12] : P0 (paradigme 0), qui correspond *grosso modo* à l'argument externe ; P1 (≈ Objet Direct) ; P2 (≈ Objet Indirect, les formes non clitiques présentant la préposition *à*).

D'autres paradigmes sont le paradigme locatif (PL) ou délocatif (PDL), le paradigme de manière (PM), le paradigme de temps (PT), ou le paradigme de quantité (PQ).

Le tableau suivant illustre l'emploi de ces paradigmes :

Verbe	type	SA	
circuler	intransitif	P0	*l'eau froide circule dans ces tubes*
disparaître	inaccusatif	P0 (PDL)	*son ami a disparu de la cabine téléphonique*
construire	transitif	P0 (P1)	*l'hiver, ils construisent des cabanes*
contribuer	transitif ind.	P0 (P2)	*son action contribue à la destruction des structures établies*
concéder	(di)transitif	P0 P1 (P2)	*on ne lui a concédé aucun droit*

Tableau 3 : illustration des paradigmes les plus courants utilisés dans la ressource DicoValence

4.2 LEC (Mel'čuk & *al*., 1995)

D'un point de vue méthodologique, en ce qui concerne les travaux de recherche existants sur la SA des nominalisations, nous nous sommes appuyés essentiellement sur les méthodes et les données des travaux initiaux de la Lexicologie Explicative et Combinatoire (Mel'čuk & *al.*, 1995) développés ensuite dans le cadre de la Théorie Sens-Texte, et notamment dans le *Lexique Actif du Français* (LAF) (Mel'čuk & Polguère, 2007)[13] .

Nous avons justement réutilisé cette modélisation des actants sémantiques en X, Y, Z pour décrire la formule actancielle et la structure argumentale des nominalisations déverbales de notre lexique et pour formaliser la codification de la réalisation syntaxique de surface de la structure argumentale des noms. Ainsi, pour ne citer ici que quelques exemples, pour le nom *création*, on trouve dans le lexique la SA « création de Y par X », pour *apparition*, la SA « apparition de X », et pour *autorisation*, « autorisation de Y à Z par X ».

4.3 Annotation du lexique

Dans le lexique, la description d'un lexème prédicatif a donc impliqué initialement :

- d'une part, l'utilisation des paradigmes de *DicoValence* en P0, P1, P2, pour décrire la SA du verbe initial dont dérive le nom déverbal donné ;
- d'autre part, l'emploi d'une forme propositionnelle dans laquelle les variables utilisées représentent les actants sémantiques (X, Y, Z, etc.) de ce lexème pour décrire la SA du nom déverbal correspondant.

Pour la SA des verbes, nous avons donc adapté *DicoValence* en transformant son système de notation (en P0, P1, P2, etc.) par celui de la *LEC* et du *LAF* (en X, Y, Z, etc.) et pour celle de noms, nous réinterprétons la SA des verbes.

Ainsi, pour le couple *proposer*/*proposition*, on avait la SA verbale « P0 proposer P1 à P2 » et la SA nominale « proposition de Y à Z par X », où X représente, dans ce cas précis, l'argument externe (le sujet) du verbe, Y le premier argument interne (le COD) du verbe transitif direct *proposer*, et Z le deuxième argument interne (le COI) du verbe.

Pour le couple *adhérer*/*adhésion*, on avait la SA verbale « P0 adhérer à P2 » et la SA nominale « adhésion de X à Y », où X représente, dans ce cas précis, l'argument externe (le sujet) du verbe et Y le premier argument interne du verbe transitif indirect *adhérer*.

Dans la notation, nous avons éliminé l'optionalité des arguments '()', notamment, parce que chez les noms, les arguments sont beaucoup plus optionnels que chez les verbes.

Nous avons également voulu représenter quelques alternances « productives » entre deux SA de la même acception ou UL, comme dans le cas de certains verbes inaccusatifs (*e.g. accélérer*) et de l'alternance anticausative (*casser / se casser*). Pour ces cas, on introduit une optionalité X V Y / X V, qui est héritée par le nom dérivé : N de Y par X / N de X.

	SA verbes	SA noms
éclore / éclosion	X éclore	l'éclosion de X
construire / construction	X construire Y	construction de Y par X
contribuer / contribution	X contribuer à Y	contribution de X à Y
proposer / proposition	X proposer Y à Z	proposition de Y à Z par X

Tableau 4 : illustration d'alternances « productives » entre deux SA de la même acception

C'est cette codification qui a été adoptée et utilisée dans le lexique pour la présentation des nominalisations de ce lexique, mais également pour la description de la réalisation syntaxique des arguments de ces nominalisations dans le corpus.

5 Quelques données

5.1 La SA du lexique

5.1.1 Patrons verbaux et nominaux

Dans le lexique, nous avons distingué 47 patrons de SA différents pour les verbes et 57 patrons de SA différents pour les noms.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

Les patrons les plus fréquents pour les verbes (plus de 15 ULs) sont ceux qui apparaissent dans le tableau 5 ci-dessous.

Comme on peut le constater, les verbes transitifs directs à deux arguments, X V Y, sont les plus fréquents (252 ULs), suivis des verbes qui entrent dans la diathèse causative, X V Y/ X V, (85 ULs) et des verbes transitifs indirects à trois arguments, X V Y à Z (41 ULs).

Patrons SA verbes	Nb ULs	Exemple
X V Y	252	*construire*
X V Y / X V	85	*améliorer*
X V Y à Z	41	*distribuer*
X V Y de Z	28	*exclure*
X V à Y	26	*adhérer*
X V	25	*éclore*
X V de Y	20	*décider*
X V Y à/dans Z	16	*délocaliser*
X V Y de Z / X V de Y	16	*augmenter*

Tableau 5 : quinze patrons de SA verbale les plus fréquents

Les patrons les plus fréquents pour les noms (plus de 15 ULs) sont ceux qui apparaissent dans le tableau 6 ci-dessous.

Patrons SA noms	Nb ULs	Nb occ.	Exemple
N de Y par X	207	776	*construction*
N de Y par X / N de X	85	361	*amélioration*
N	69	470	*administration*
N de X	30	125	*éclosion*
N de X à Y	28	150	*adhésion*
N de Y à Z par X	28	136	*distribution*
N de Y de Z par X	27	134	*exclusion*
N de X de Y	20	305	*décision*
N de X à/dans Y	17	162	*circulation*
N de Y de Z par X / N de X de Y	17	97	*augmentation*

Tableau 6 : quinze patrons de SA nominale les plus fréquents

Un cas particulier est celui des noms concrets, soit parce qu'ils désignent un résultat, comme la *construction* (*e.g. cette construction en pierre*) soit parce qu'ils désignent un collectif (*administration*,

rédaction). Malgré le fait que, strictement parlant, ces noms n'ont pas d'arguments, il est pertinent de les inclure ici, comme des noms ayant une SA vide ({ø}), parce qu'ils permettent de poser des questions intéressantes. Par exemple, est-ce qu'ils dérivent directement des verbes ou dérivent-ils plutôt d'une autre acception du même nom ?

Un autre cas intéressant est celui des patrons comme N de X à Y (*adhésion*) ou N de X de Y (*décision*). Normalement, dans la SA des noms dérivés de verbes transitifs, l'argument correspondant à l'argument verbal externe peut être récupéré moyennant une sorte de complément d'agent nominal (introduit par *par*), comme dans *le bombardement de la ville par les anglais*.

(2) a. $[\text{Le bombardement}]_N$ de $[\text{la ville}]_Y$ par $[\text{les anglais}]_X$

b. $[\text{Les anglais}]_X$ $[\text{ont bombardé}]_V$ $[\text{la ville}]_Y$

Les noms du type de *adhésion* ou *décision* ne suivent pas ce modèle, selon lequel l'argument correspondant à l'OD du verbe est réalisé comme le premier argument de la nominalisation, mais cet autre :

(3) a. $[\text{L'adhésion}]_N$ de $[\text{Dominique}]_X$ au $[\text{Parti Socialiste}]_Y$

b. $[\text{Dominique}]_X$ $[\text{a adhéré}]_V$ au $[\text{Parti Socialiste}]_Y$

(4) a. $[\text{La décision}]_N$ de $[\text{Pierre}]_X$ de $[\text{partir}]_Y$

b. $[\text{Pierre}]_X$ $[\text{a décidé}]_V$ de $[\text{partir}]_Y$

Le comportement des noms du type de *adhésion* et *décision* n'est pas dû à une différence dans le rôle thématique du sujet verbal, parce que dans tous les cas il s'agit d'un agent ; il faudrait donc plutôt aller chercher l'explication du côté de l'argument verbal interne.

Mais dans tous les cas, ces exemples mettent en évidence l'intérêt d'une recherche exhaustive à mener à partir des données réelles.

5.1.2 De la SA verbale à la nominale

Si l'on compare la SA des noms et celle des verbes dans le lexique, on peut constater clairement que l'on trouve très rarement des cas de décalage, c'est-à-dire, des cas où la SA nominale ne suit pas de façon fidèle celle du verbe.

Comme l'illustrent les exemples mentionnés dans le tableau suivant, les cas de décalage constituent à peine 7% (24 cas) du total des cas (339).

SA des noms	SA des verbes	Exemple	Calques	Décalages
N de X	X V	éclosion > éclore	21	9
N de Y par X	X V Y	construction > construire	204	3
N de X à Y	X V à Y	adhérer > adhésion	21	7
N de X de Y	X V de Y	décider > décision	14	6
N de Y à Z par X	X V Y à Z	distribuer > distribution	28	0
N de Y de Z par X	X V Y de Z	exclure > exclusion	27	0
			315	24

Tableau 7 : cas de calques et de décalages de la SA du déverbal et de la SA du verbe de base

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

Parmi les cas de décalage entre la SA du verbe et celle du nom, on peut distinguer deux cas de figure très majoritaires :

- perte d'un argument : l'*invention* de Paul ← X *inventer* Y
- changement de préposition : *avertissement* de X à Y ← X *avertir* Y

Mais ce qu'il convient de souligner ici, c'est que, dans notre lexique, on ne trouve pas de noms incluant un argument qui n'était pas déjà présent dans le verbe de base. Ces résultats respectent donc notre hypothèse de départ sur la préservation de la SA (HPSA).

Nous allons maintenant analyser les résultats que l'on obtient en comparant cette SA idéale des noms avec sa réalisation concrète en corpus.

5.2 Comparaison lexique-corpus

Comme nous l'avons déjà signalé, notre lexique est construit à partir d'un corpus, de sorte que chaque acception est liée à ses occurrences respectives dans le corpus. Le *ratio* entre les occurrences (4018) et les acceptions (742) est de 5,4 en moyenne.

Hormis très peu d'exceptions, on peut dire de façon générale que dans la SA des noms, X correspond toujours à l'argument externe du verbe, tandis que Y correspond généralement à un argument interne (pas toujours un OD)[14]. De ce fait, nous allons pouvoir mettre en relation, en dernière instance, la réalisation de la SA des noms en corpus avec celle des verbes d'origine.

Les données que nous allons analyser sont résumées dans le tableau 8 ci-dessous, où nous avons inclus les quatre patrons les plus fréquents, hormis le patron N de Y par X / N de X[15] correspondant aux noms dérivés des verbes qui entrent dans l'alternance anticausative.

	ø	Y	X	X et Y	Total
N de Y par X	217	394	13	12	636
	34,1%	61,9%	2%	1,9%	
N de X de Y	126	9	72	1	208
	60,6%	4,3%	35%	0,5%	
N de X à Y	63	15	18	7	103
	61,2%	14,6%	17,5%	6,8%	
N de X	86		29		115
	74,8%		25,2%		

Tableau 8 : quatre patrons de SA nominale les plus fréquents

Ce qui attire d'abord l'attention, c'est que, pour les quatre patrons analysés, les cas où les deux arguments (X et Y) apparaissent effectivement réalisés dans le corpus sont vraiment peu nombreux : sauf pour le patron *N de X à Y* (6,8%), pour les autres patrons, le pourcentage n'arrive même pas à 2%.

Si on part du pôle opposé, c'est-à-dire, les cas du corpus où aucun des deux arguments n'apparaît, les résultats sont encore plus intéressants : sauf pour le patron *N de Y par X*, les cas d'emploi absolu des noms sont les plus fréquents, et presque toujours supérieurs à 50% : 60,6% pour *N de X de Y* ; 61,2% pour *N de X à Y*, et 74,8% pour *N de X*.

Pour le patron N de Y par X, l'emploi absolu est de 34,1%, la réalisation de Y étant le cas de réalisation le plus fréquent (61,9%). Pour le reste des patrons, la réalisation de Y est beaucoup moins fréquente, car elle arrive en effet à peine à 20% : 19,3% (*N de Y par / N de X*) ; 14,6% (*N de X à Y*), et 4,3% (*N de X de Y*). Ces résultats ne vont pas tout à fait dans le sens de la prédiction de Grimshaw (1990), pour qui la réalisation de Y (l'argument chez les noms correspondant à l'argument interne des verbes d'origine) devrait être (presque) obligatoire.

Enfin, la réalisation de X arrive généralement à peine à un tiers des cas : 2% (N de Y par X) ; 17,5% (N de X à Y) ; 35% (N de X de Y), et 25,2% (N de X).

5.3 La SA des occurrences des déverbaux en corpus

5.3.1 L'annotation du corpus

Dans le corpus, on trouve la description systématique de la SAS des noms déverbaux. Il s'agit d'une description de la réalisation syntaxique des arguments sémantiques des noms prédicatifs déverbaux en corpus (et pas celle des verbes dont ils sont dérivés, non décrits en corpus) : celle-ci n'est pas effectuée au niveau du lexique mais à celui du corpus, c'est-à-dire au niveau des occurrences effectives des nominalisations déverbales dans le corpus de travail.

Cette description comprend deux informations :

- si un argument est réalisé ou non ;
- si oui, comment il est réalisé.

Pour l'encodage de la réalisation syntaxique des arguments, c'est-à-dire le codage du type de complément du nom déverbal, qui s'inspire également des descriptions produites dans le cadre de la LEC (Mel'čuk & *al.*, 1995), nous avons choisi de distinguer :

- les groupes nominaux compléments employés avec déterminant, encodés comme Gdét (ex : *la construction d'un logement*), qui impliquent la présence d'un article (défini ou indéfini, singulier ou pluriel) ;
- les groupes nominaux compléments employés sans déterminant, encodés comme GN (ex : *la construction de logement*), au singulier ou au pluriel, sans article apparent dans la réalisation syntaxique de surface.

Cela permet d'établir ainsi une distinction syntaxique qui s'appuie sur les constituants de la grammaire traditionnelle, entre :

- « la construction du logement », où « du logement », mis pour « de + le logement », = « de Gdét (défini au singulier) » ;
- « la construction des logements », où « des logements », mis pour « de + les logements » = « de Gdét (défini au pluriel) » ;
- « la construction d'une maison », où « d'une maison » = « de Gdét (indéfini au singulier) » ;
- « la construction de logements », où « de logements », mis dans la réalisation de surface pour « de + des logements » dans la structure syntaxique profonde, = « de GN (au pluriel) » ;
- « la construction de logement », où « de logement » = « de GN (au singulier) ».

On a également pris en compte la codification syntaxique des arguments « cachés », tels que :

- l'adjectif possessif, abrégé en « det poss » (ex. : *son adhésion au FMI*) ;
- l'adjectif relationnel, abrégé en « adj rel » (ex. : *l'adhésion russe au FMI*).

Le tableau suivant inclut quelques exemples :

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

GDét	*la construction du / d'un logement*
GN	*la construction de logement(s)*
det poss	*son adhésion au FMI*
adj rel	*l'adhésion russe au FMI*
nom propre	*la décision de Nicolas*
Vinf	*la décision d'investir dans l'immobilier*

Tableau 9 : exemples des étiquettes pour l'annotation du corpus

Dans notre lexique, il est donc également possible de vérifier quelle est la réalisation syntaxique des différents arguments.

PATRON	X	Nb d'occurrences
N de X	Ø	86
N de X	de Gdet	29
N de X	de GN	5
N de X	adj rel	2
N de X	det poss	2

Tableau 10 : réalisation syntaxique de l'argument pour le patron « N de X »

PATRON	X	Y	Nb d'occurrences
N de Y par X	Ø	de Gdet	317
N de Y par X	Ø	Ø	255
N de Y par X	Ø	de GN	78
N de Y par X	Ø	adj rel	18
N de Y par X	Ø	det poss	13
N de Y par X	de Gdet	Ø	11
N de Y par X	par Gdet	de Gdet	11
N de Y par X	det poss	Ø	10

Tableau 11 : réalisation syntaxique des arguments pour le patron « N de Y par X »

Comme on peut le voir dans les tableaux 10 et 11, la réalisation syntaxique qu'adoptent le plus fréquemment les arguments est Gdet (groupe nominal employé avec déterminant).

6 Perspectives

La ressource que nous venons de présenter permet, grâce aux informations contenues sur la structure argumentale des noms déverbaux et des bases verbales correspondantes, de vérifier les postulats théoriques suivants. Notre hypothèse de la préservation de la SA (HPSA) a été vérifiée : la SA des noms déverbaux, dans la majorité des cas, est héritée de la SA du verbe correspondant. Nous avons montré que l'argument qui correspond à l'OD du verbe n'est pas systématiquement obligatoire chez les nominalisations (Grimshaw, 1990) démontrant ainsi que les arguments de la structure argumentale des noms déverbaux sont plus optionnels que les arguments de la structure argumentale des verbes correspondants (Péris & Taulé, à par.).

Concernant l'amélioration et le suivi des données relatives à l'étude de la structure argumentale des noms déverbaux, il est envisageable, entre autre, de réaliser le typage sémantique des actants voire leur étiquetage en rôles thématiques.

Concernant les perspectives de recherche qu'offre notre ressource, nous envisageons d'exploiter les données concernant la classe aspectuelle des noms déverbaux et de leur verbe correspondant. Il s'agira alors de vérifier l'hypothèse selon laquelle la classe aspectuelle du nom déverbal est héritée de celle du verbe. Cette étude de la classe aspectuelle pourra être croisée avec celle de la structure argumentale et constituer les prémices de la reconnaissance automatique en corpus de la classe aspectuelle.

Enfin, concernant la consultation et l'exploitabilité des données de notre ressource, nous avons réalisé une interface de consultation accessible en ligne (http://nomage.recherche.univ-lille3.fr/nomage/) ainsi qu'un lexique adapté selon la norme Lexical Markup Framework (LMF).

Références

Abeillé, A. (2003). *Treebanks, Building and Using Parsed Corpora*, Dordrecht : Kluwer.

Balvet, A., L. Barque, M-H. Condette, P. Haas, R. Huyghe, R. Marín & A. Merlo (2011). *Proceedings of the First International Workshop on Lexical Recources (WoLeR 2011),* pp. 8-15.

Eynde, K. van den & P. Mertens (2006). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13/1, pp. 63-104.

Gardent, C., B. Guillaume, G. Perrier & I. Falk (2006). *Extraction d'information de sous-catégorisation à partir des tables du LADL*. Actes TALN 2006.

Gardent, C. (2009). Evaluating an automatically extracted lexicon. *4th Language & Technology Conference, Poznan, Poland.*

Grimshaw, J. (1990). *Argument Structure*. Cambridge : MIT Press.

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

Kupść A. & A. Abeillé (2008). Growing TreeLex. In Gelbukh A. (éd.), *9th Int. Conf., CICLing 2008*, (Haifa, Israel, February 2008), Lecture Notes in Computational Linguistics, 4919, pp. 28-39.

Leclère, C. (2002). Organisation of the Lexicon-Grammar of French Verbs. *Linguisticæ Investigationes*, 1, vol. 25.

Macleod, C., A. Meyers, R. Grishman, L. Barrett & R. Reeves (1998). NOMLEX: A Lexicon of Nominalizations. *Proceedings of EURALEX'98, Liège, août 1998*.

Mel'čuk, I.A., A. Clas, & A. Polguère (1995). *Introduction à la lexicologie explicative et combinatoire*. Paris: Duculot.

Mel'čuk, I.A. & A. Polguère (2007). *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Bruxelles : Éditions De Boeck.

Messiant, C., A. Korhonen & T. Poibeau (2008). Lexschem : A large subcategorization lexicon for french verbs. *Language Resources and Evaluation Conference (LREC), Marrakech*.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100271

Meyers A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young & R. Grishman (2004). Annotating Noun Argument Structure for NomBank. *Proceedings of LREC-2004.*

Peris, A. & M. Taulé (à par.). Annotating the Argument Structure of Deverbal Nominalizations in Spanish, *Language Resources and Evaluation.*

Sagot, B., L. Clément, E. Villemonte de la Clergerie & P. Boullier (2006). The Lefff 2 syntactic lexicon for French : architecture, acquisition, use. *Actes de LREC 06, Gênes, Italie.*

Salkoff M. & A. Valli (2005). A dictionary of french verbal complementation. 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.

Taulé, M., M. A. Marti & M. Recasens (2008). Ancora: Multilevel Annotated Corpora for Catalan and Spanish. Proceedings of 6th International Conference on Language Resources and Evaluation, Marrakesh (Morocco).

Tolone, E. (2011). Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français. Thèse de doctorat, LIGM, Université Paris-Est, France.

---

[1] Les tables du Lexique-Grammaire sont librement téléchargeables sur http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Telechargement.html

[2] Corpus journalistique constitué de 10 années d'articles du journal *Le Monde.*

[3] Dans les travaux de Korhonen (2002) notamment, la sous-catégorisation est utilisée pour décrire également des phénomènes sémantiques.

[4] Le lexique LexSchem est librement téléchargeable et consultable sur http://www-lipn.univ-paris13.fr/~messiant/lexschem.html

[5] TreeLex est téléchargeable sur http://erssab.u-bordeaux3.fr/spip.php?article150

[6] Cf. sur *Nomlex* : http://nlp.cs.nyu.edu/nomlex/index.html

[7] Cf. site sur *Nombank* : http://nlp.cs.nyu.edu/meyers/NomBank.html et l'article de Adam MEYERS, Ruth REEVES, Catherine MACLEOD, Rachel SZEKELY, Veronika ZIELINSKA, Brian YOUNG et Ralph GRISHMAN. *The NomBank Project: An Interim Report*. New York University.

Document consultable et téléchargeable en format PDF sur : http://nlp.cs.nyu.edu/meyers/papers/nombank-pap.pdf

[8] *PropBank*, pour Proposition Bank, est un projet d'annotation sémantique de corpus de texte élaboré par l'Université de Colorado aux Etats-Unis, sous l'égide de Martha Palmer. L'objectif du projet *PropBank* est d'étiqueter les structures argumentales prédicatives dans le *Penn Treebank*, avec des étiquettes relatives au sens et des étiquettes relatives aux arguments qui sont basées sur les entrées lexicales du *Penn Treebank* contenant des informations relatives à leurs structures argumentales prédicatives, les relations entre prédicat et argument ayant été ajoutées aux arbres syntaxiques du *Penn Treebank*.

[9] *cf.* Présentation générale du projet *DicoValence* consultable sur Internet à l'adresse :

http://bach.arts.kuleuven.be/DicoValence/

[10] Dictionnaire de valence au format texte .txt disponible sous différents encodages (ISO-8859-1 pour Windows, UTF-8 pour Mac et Linux, compressé au format Zip) consultable et téléchargeable à l'adresse :

http://bach.arts.kuleuven.be/DicoValence/#fichiersfr

[11] Karen van den EYNDE et Piet MERTENS. *Le dictionnaire de valence DicoValence : manuel d'utilisation*, version 1.2, p. 2. Document consultable et téléchargeable sur : http://bach.arts.kuleuven.be/DicoValence/manuel_061117.pdf

[12] *ibid*., p. 5.

[13] *cf.* site Internet du *LAF* : http://olst.ling.umontreal.ca/laf/le-laf/

[14] La seule exception régulière correspond aux verbes psychologiques à expérienceur objet du type de *préoccuper*.

[15] Nous n'avons pas tenu compte de ce patron parce qu'il ne nous permet pas de tirer des conclusions sur la distinction entre les arguments X et Y.