

UNIVERSITÉ DE LA MÉDITERRANÉE  
U.F.R. SCIENCES DE LUMINY

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE, E.D. 184

**THÈSE**

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE LA MÉDITERRANÉE

*Spécialité : Informatique*

par

**Mohamed Mahdi MALIK**

sous la direction de Paul Sabatier et Jean Royauté

*Titre :*

**STRUCTURES PRÉDICATIVES NOMINALES EN ANGLAIS :  
ACQUISITION DE DONNÉES LEXICALES  
POUR L'ANALYSE AUTOMATIQUE DE TEXTES**

soutenue publiquement le 28 janvier 2010

JURY

Mme. Brigitte GRAU	Professeur, LIMSI, ENSIIE, Orsay	<i>Rapporteur</i>
M. Guy LAPALME	Professeur, RALI, Université de Montréal	<i>Rapporteur</i>
M. Alexis NASR	Professeur, Université de la Méditerranée	<i>Examineur</i>
M. Jean ROYAUTÉ	Chargé de Recherche au CNRS, Université de la Méditerranée	<i>Directeur</i>
M. Paul SABATIER	Directeur de Recherche au CNRS, Université de la Méditerranée	<i>Directeur</i>
M. Max SILBERZTEIN	Professeur, Université de Franche-Comté	<i>Examineur</i>



# Table des matières

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Ressources électroniques verbales et déverbales</b>	<b>9</b>
2.1	Introduction . . . . .	10
2.2	Les ressources lexico-syntaxiques . . . . .	10
2.2.1	Les lexiques COMLEX . . . . .	11
2.2.2	De COMLEX à NOMLEX . . . . .	18
2.2.3	The SPECIALIST Lexicon . . . . .	22
2.3	Les ressources syntaxico-sémantiques . . . . .	25
2.3.1	WordNet . . . . .	26
2.3.2	FrameNet . . . . .	28
2.3.3	VerbNet . . . . .	36
2.3.4	PropBank . . . . .	42
2.3.5	NomBank . . . . .	43
2.4	Conclusion . . . . .	46
<b>3</b>	<b>Structures Prédicatives</b>	<b>51</b>
3.1	Introduction . . . . .	52
3.2	Les structures argumentales prédicatives . . . . .	53
3.3	Les prédicats . . . . .	54
3.3.1	L'acquisition Automatique de Cadres de Sous-Catégorisation . . . . .	54
3.3.2	L'acquisition automatique des contraintes de sélection . . . . .	57
3.3.3	L'acquisition d'informations sémantiques sur les aspects verbaux . . . . .	58
3.3.4	La classification sémantique des verbes . . . . .	59
3.3.5	La compréhension de textes multilingues . . . . .	60
3.4	Les liens lexicaux entre les différents prédicats . . . . .	60
3.4.1	Les liens morphologiques . . . . .	60

3.4.2	Les liens morpho-sémantiques . . . . .	65
3.4.3	Conclusion . . . . .	68
<b>4</b>	<b>Acquisition des Structures Argumentales</b>	<b>69</b>
4.1	Introduction . . . . .	70
4.2	Le groupe nominal . . . . .	70
4.2.1	Structure distributionnelle du groupe nominal . . . . .	70
4.2.2	Le groupe nominal à tête prédicative . . . . .	75
4.2.3	Les formes équivalentes du groupe nominal prédicatif . . . . .	77
4.2.4	Les prépositions, marqueurs argumentaux du GNpréd . . . . .	80
4.3	Une base de données pour l'acquisition des prédicats nominaux . . . . .	83
4.3.1	L'architecture générale de PredicateDB . . . . .	83
4.3.2	Les différents modules de PredicateDB . . . . .	85
4.3.3	Détermination des types argumentaux . . . . .	89
4.4	La création d'un lexique de nominalisations . . . . .	95
4.4.1	Classification des nominalisations . . . . .	95
4.4.2	Les ambiguïtés liées aux super-classes . . . . .	105
4.4.3	Constitution du lexique de nominalisations . . . . .	112
4.4.4	Traitement des entrées lexicales complexes . . . . .	112
4.4.5	Les propriétés non traitées . . . . .	116
4.5	Le lexique PredicateDB pour l'acquisition de structures argumentales . . . . .	119
4.5.1	Le Link Parser et les grammaires de liens . . . . .	119
4.5.2	Une grammaire pour les prédicats nominaux . . . . .	121
4.5.3	Analyse des GNpréd avec le Link Parser . . . . .	124
4.5.4	Pertinence du lexique et des analyses . . . . .	127
4.6	Conclusion . . . . .	128
<b>5</b>	<b>Conclusion</b>	<b>130</b>

# Chapitre 1

## Introduction

Notre travail est consacré aux structures prédicatives et tout particulièrement à la relation singulière qui unit un prédicat verbal et un prédicat nominal. Ces deux types de prédicats possèdent des structures argumentales identiques mettant en jeu des informations communes (exemples 1.1a et 1.1d). Les prédicats nominaux dotés de leurs arguments donnent lieu à des groupes nominaux prédicatifs (GNpréd). La mise en évidence de l'équivalence des structures argumentales des prédicats verbaux et des GNpréd est fondamentale pour pouvoir réaliser, par exemple, des systèmes d'extraction automatique d'informations très performants en Traitement Automatique des Langues (TAL).

Nous nous intéressons aux textes scientifiques et tout particulièrement aux textes de biologie et médecine. Il s'agit d'un domaine où les travaux en fouille de textes, terme générique qui désigne la plupart des travaux de TAL, sont particulièrement importants. Afin de rendre plus performantes les applications, une partie de ces travaux repose sur l'utilisation de ressources linguistiques électroniques, telles que COMLEX [Grishman et al., 1994], PropBank [Palmer et al., 2005] et NOMLEX [Macleod et al., 1998], qui décrivent les caractéristiques morphologiques, syntaxiques et sémantiques de l'anglais ainsi que d'autres informations qui sont plus complexes telles que le comportement de certaines unités lexicales dans différents contextes syntaxiques, les structures syntaxiques de leurs compléments, leurs relations lexico-syntaxiques.

Les propriétés des structures prédicatives peuvent être utilisées dans la capture des régularités syntaxiques et fournir ainsi une représentation commune à partir des différentes variantes syntaxiques. Ces dernières sont dues au fait que le langage naturel dispose de plusieurs réalisations syntaxiques pour exprimer le même sens. Une phrase peut être écrite par exemple sous la forme active (exemple 1.1a), passive (exemple 1.1b), infinitive (exemple 1.1c), gérondive

(exemple 1.1d) ou encore sous la forme d'un groupe nominal prédicatif dans lequel le prédicat verbal est remplacé par une nominalisation (exemple 1.1e).

- (1.1) (a) *activation of macrophages expresses cytotoxic activity* . . .  
(l'activation des macrophages exprime l'activité cytotoxique . . .)
- (b) *cytotoxic activity is expressed by the activation of macrophages* . . .  
(l'activité cytotoxique est exprimée par l'activation des macrophages . . .)
- (c) *activation of macrophages to express cytotoxic activity was* . . .  
(l'activation des macrophages pour exprimer l'activité cytotoxique était . . .)
- (d) *activation of macrophages expressing cytotoxic activity was* . . .  
(l'activation des macrophages exprimant l'activité cytotoxique était . . .)
- (e) *the expression of cytotoxic activity by the activation of macrophages* . . .  
(l'expression de l'activité cytotoxique par l'activation des macrophages . . .)

La plupart des travaux portant sur la reconnaissance des structures prédicatives se sont davantage intéressés aux schémas verbaux (sujet, verbe et compléments) car une analyse syntaxique simple permet d'accéder naturellement aux prédicats verbaux et à leurs arguments. En règle générale, dans ce type de schémas, l'argument qui précède le prédicat est considéré comme un sujet et celui ou ceux qui le suivent sont des compléments. Par contre, les arguments nominaux d'un GNpréd présentent plusieurs difficultés d'analyses. Ils ont la particularité d'être mobiles, ils peuvent être placés avant le prédicat nominal en position prémodifieur (nom, possessif, déterminant) ou en position postmodifieur, situés après le prédicat et introduits dans ce cas par des prépositions spécifiques. Les GNpréd se caractérisent aussi par le fait que tous leurs arguments sont optionnels et peuvent ainsi être effacés. Ces difficultés, inhérentes à leur traitement automatique, font que les recherches en TAL se focalisent davantage sur les schémas verbaux. Cependant, le fait que la structure argumentale des prédicats nominaux ne soit pas analysée a pour conséquence une perte d'informations en extraction d'informations, dans un contexte où les documents scientifiques et surtout ceux qui appartiennent au domaine de la biologie en font un usage abondant. En effet, il est fréquent d'utiliser le GNpréd : *VEGF regulation by Heregulin* (la régulation du VEGF par l'Heregulin) à la place de la phrase qui lui correspond : *Heregulin regulates VEGF* (l'Heregulin régule le VEGF).

L'analyse des structures prédicatives nominales requière la mise en évidence des relations argumentales ou des fonctions syntaxiques des différents arguments appartenant aux

GNpréd (sujet, complément d'objet direct, complément prépositionnel). Les positions des arguments verbaux que l'on retrouve dans le GNpréd dépendent de la classe grammaticale du prédicat verbal qui lui est lié (intransitif, transitif, ditransitif) ainsi que du type des arguments verbaux (humain, objets). Plusieurs méthodes ont été développées dans le but de déduire ces liens. La plupart d'entre elles partent du type de la classe verbale où les propriétés des verbes sont connues pour déduire toutes les combinaisons possibles des arguments relatifs aux nominalisations dérivées de ces verbes [Nunes, 1993]. A l'inverse, d'autres méthodes [Gurevich et al., 2006] prennent comme point de départ la structure du groupe nominal où les arguments nominaux sont déjà nommés, et la tâche consiste à les relier aux cadres de sous-catégorisation verbales. En s'appuyant sur les principes de la première méthode, nous formalisons les conditions dans lesquelles se réalisent les relations d'équivalence entre les deux types de constructions (verbales et nominales), à partir d'un lexique anglais de biologie, le Specialist Lexicon [Browne et al., 2000]. En nous basant sur les données de ce lexique, que l'on peut qualifier de génériques car elles couvrent un vaste domaine large qui va de la biologie et de la médecine aux sciences sociales, nous confirmons le rôle que peuvent jouer les prépositions dans la détermination et l'acquisition des différents arguments nominaux. Nous mettons en évidence les structures distributionnelles et fonctionnelles du GNpréd ainsi que les différentes formes équivalentes qu'il peut avoir. Nous montrons que le comportement syntaxique des GNpréd dépend de la classe grammaticale du prédicat verbal à partir duquel la nominalisation a été dérivée et de son schéma de complémentation. Les prédicats verbaux peuvent être répartis en deux grandes classes : les prédicats transitifs qui admettent un complément d'objet direct (COD) et les prédicats qui n'en admettent pas. Les GNpréd associés à ces deux classes se caractérisent par le fait que la préposition *of* sélectionne soit le sujet, soit le COD et qu'elle est un marqueur ambigu lorsqu'elle apparaît seule. Cependant, cette ambiguïté disparaît lorsque les GNpréd sont saturés, c-à-d, lorsque tous les arguments de la relation décrite par le nom prédicatif sont réalisés.

Cette étude nous a conduit à développer la plate-forme PredicateDB [Malik and Royauté, 2009, 2007] qui, à partir des données extraites du Specialist Lexicon, nous permet de définir pour l'anglais les relations syntaxiques qui lient les schémas verbaux à leurs correspondants nominaux. Notre travail a consisté à donner une prédiction raisonnable du comportement syntaxique des arguments nominaux lorsqu'ils sont en position postmodifieur, en confirmant le rôle de marqueur que peuvent jouer les prépositions dans la détermination et l'acquisition des différents arguments nominaux (sujet, complément d'objet direct, complément d'objet indirect). Cela nous a permis ensuite d'associer à chaque nominalisation appartenant au Specialist Lexicon, le ou les GNpréd qui lui correspondent. A partir de ces informations, PredicateDB nous a permis de créer semi-automatiquement un lexique dont le rôle est de

décrire le comportement syntaxique de toutes les nominalisations qui appartiennent au Specialist Lexicon. Ce lexique a été utilisé dans le développement d'une grammaire de liens permettant d'identifier les arguments des nominalisations de la même façon que pour un verbe. Cette grammaire a été intégrée dans la grammaire du Link Parser [Sleator and Temperley, 1991] dans le but d'analyser syntaxiquement des GNpréd et d'acquérir leurs arguments nominaux à partir de textes de biologie [Royauté et al., 2006, 2007, Godbert and Royauté, 2009].

Cette thèse est composée de trois parties : le premier chapitre présente un état de l'art dans lequel on effectue une étude des différentes ressources électroniques (lexiques et corpus) qui existent et celles qui peuvent être intéressantes pour notre étude. Parmi ces ressources, nous distinguons : (i) les ressources lexico-syntaxiques qui décrivent les propriétés lexicales et/ou syntaxiques qui caractérisent les différentes unités lexicales, (ii) les ressources syntaxico-sémantiques qui détaillent les propriétés syntaxiques et/ou sémantiques des différents items ainsi que les relations qui les lient. Le deuxième chapitre est une étude détaillée des structures argumentales prédictives (SAP) et de leurs propriétés syntaxiques et sémantiques. On montre leur importance dans la capture des régularisations syntaxiques en permettant une représentation commune des différentes formes syntaxiques (forme active, forme passive, forme avec mouvement du datif). Dans ce chapitre nous faisons référence aux différents travaux qui ont eu recours à cette structure. Nous nous sommes également intéressé aux travaux qui portent sur les liens qui existent entre les différents prédicats (liens morphologiques et morpho-syntaxiques) et qui permettent de déduire les propriétés communes que partagent les différents prédicats. Nous examinons également comment ces différents types de liens ont été utilisés pour construire des lexiques comme WordNet [Miller, 1985, 1995, Fellbaum, 1998], HowNet [Dong, 2000], CatVar [Habash and Dorr, 2003]. Le troisième chapitre décrit la structure distributionnelle du groupe nominal, qu'il soit prédictif ou non prédictif, et les différences fonctionnelles qui caractérisent les constituants de chaque groupe nominal. Dans ce chapitre nous décrivons la plate-forme PredicateDB : son architecture, les différents modules qui la composent, ainsi que l'utilisation que nous en avons faite pour déterminer les types argumentaux pour la création d'un lexique de nominalisations.

## **Chapitre 2**

# **Ressources électroniques verbales et déverbales**

## 2.1 Introduction

Les ressources électroniques ont connu ces dernières années un regain d'intérêt chez les chercheurs travaillant dans le domaine du traitement automatique du langage durant ces dernières années. Cet intérêt est dû principalement à leur utilisation dans les différentes applications destinées au TAL. Nous distinguons des ressources lexico-syntaxiques dont le principal rôle est de mettre en évidence des propriétés lexicales et/ou syntaxiques qui caractérisent les différentes unités lexicales (noms, verbes, adjectifs, adverbes, etc.). D'autres types de ressources, que nous nommons ressources syntaxico-sémantiques, s'intéressent aux propriétés syntaxiques et/ou sémantiques de ces mêmes unités lexicales, ainsi qu'à leurs relations. Grâce à ces différentes données on peut, par exemple, décrire le comportement syntaxique et sémantique des prédicats nominaux et verbaux. La plupart des ressources syntaxico-sémantiques ne décrivent que la structure argumentale des prédicats verbaux et ne s'intéressent que très rarement aux nominalisations (noms déverbaux) ainsi qu'à leurs structures argumentales. Or les nominalisations sont très fréquentes dans l'anglais parlé et écrit, tout particulièrement dans les textes scientifiques. L'analyse de 2 000 phrases appartenant au Wall Street Journal a montré que plus de la moitié d'entre elles contenaient au moins une nominalisation [Gurevich et al., 2006]. Ceci montre de façon évidente que l'étude des propriétés relatives aux nominalisations est nécessaire pour le développement ou l'amélioration des différents systèmes de TAL : recherche documentaire, extraction d'informations, résumés de textes, questions réponses, traduction automatique, etc.

Dans ce chapitre, nous allons présenter différents types de ressources électroniques et détailler leurs caractéristiques. Dans la section 1, on parlera des principales ressources lexico-syntaxiques qui ont été créées. Dans la section 2, on présentera les ressources syntaxico-sémantiques. Finalement, dans la section 3, on verra un autre type de ressources syntaxico-sémantiques que sont les corpus électroniques.

## 2.2 Les ressources lexico-syntaxiques

Les ressources lexico-syntaxiques traitent essentiellement tout ce qui touche au comportement syntaxique des différents prédicats (verbaux, nominaux, adjectivaux) et plus précisément, elles étudient le comportement syntaxique des syntagmes nominaux et verbaux associés à ces prédicats. La plupart de ces ressources sont dédiées plus particulièrement à la description des prédicats verbaux. COMLEX [Grishman et al., 1994] fait partie des ressources qui mettent en évidence les propriétés syntaxiques des différents prédicats verbaux ainsi que le comportement syntaxique des compléments associés à ces prédicats. Néanmoins, on peut trouver certaines

ressources qui s'intéressent à d'autres types de prédicats tels que les nominalisations verbales et adjectivales. Parmi ces ressources on peut citer The Specialist Lexicon [Browne et al., 2000] et COMLEX-PLUS [Meyers et al., 2004a]. La première est un lexique qui contient environ 250 000 entrées appartenant à différentes catégories syntaxiques (verbes, noms, adjectifs, adverbes). Ce lexique traite non seulement les mots issus de la langue courante mais aussi les mots qui appartiennent au domaine de la biologie (*phenylmercapturic acid*, *myofascial*, *cardiovert*, etc.). La seconde ressource, COMLEX-PLUS, est un lexique qui reprend les données verbales incluses dans COMLEX et les enrichit avec les nominalisations qui leur correspondent, ainsi qu'avec certaines propriétés qui caractérisent les schémas syntaxiques des structures argumentales de ces nominalisations. D'autres types de ressources, qui sont moins nombreuses que celles qui sont dédiées aux prédicats verbaux, s'intéressent exclusivement aux prédicats nominaux (nominalisations). NOMLEX [Macleod et al., 1998] est un lexique qui fait partie de ce type de ressources. Il décrit les arguments nominaux qui sont associés à ces nominalisations et met en avant les valeurs prises par les différents arguments verbaux quand ces derniers se trouvent dans les syntagmes nominaux correspondants.

### 2.2.1 Les lexiques COMLEX

COMLEX et COMLEX-PLUS sont deux lexiques qui ont été créés dans le cadre du projet COMLEX Syntax. Ce dernier est un projet élaboré au sein de l'Université de New York en collaboration avec le LDC (Linguistic Data Consortium<sup>1</sup>). L'objectif initial de ce projet était de créer un lexique partageable et de large couverture, avec comme but la description des propriétés syntaxiques des différents mots de l'anglais afin de les exploiter pour l'analyse automatique. Le premier résultat de ce projet a été la création du lexique COMLEX [Grishman et al., 1994]. La même équipe a ensuite enrichi COMLEX avec de nouvelles caractéristiques et a utilisé ce résultat pour créer COMLEX-PLUS.

**1) COMLEX :** COMLEX est un lexique à large couverture contenant les propriétés syntaxiques d'environ 38 000 mots en anglais. La première version de ce lexique a été développée en mai 1994. Il propose des informations détaillées sur les structures des compléments verbaux et des noms et adjectifs qui admettent des compléments. Afin que les sous-catégorisations soient les plus détaillées possible, les auteurs ont exploité toutes les informations utiles figurant dans d'autres grands lexiques : le lexique verbal de Brandeis [Grimshaw and Jackendoff, 1981], le

---

1. Le Linguistic Data Consortium (Consortium Linguistique des données) est une organisation à but non lucratif qui fait partie de l'université de Pennsylvanie. Son rôle est de réunir une grande variété de données et de les mettre à la disposition des linguistes et des chercheurs en TAL

lexique créé dans le cadre des projets ACQUILEX<sup>2</sup> [Sanfilippo, 1992], celui du NYU Linguistic String Project<sup>3</sup> [Sager, 1981], ainsi que les données de deux grands dictionnaires : le dictionnaire OALD (Oxford Advanced Learner's Dictionary) et le dictionnaire LDOCE (Longman Dictionary Of Contemporary English) [Procter, 1978]. Les notations données aux différents types de compléments sont basées sur les conventions utilisées dans le lexique verbal de Brandeis où les propriétés de chaque entrée sont représentées d'une façon imbriquée et parenthésée inspirée du langage de programmation Lisp.

Les concepteurs de COMLEX ont créé manuellement leur lexique et n'ont pas eu recours aux différentes méthodes automatiques qui ont été développées durant les années 1990 et qui servaient à identifier automatiquement certaines sous-catégorisations [Brent, 1993, Manning, 1993]. Les concepteurs, qui n'ignoraient pas ces travaux, ont préféré s'en tenir aux méthodes manuelles car ils ont jugé que les méthodes automatiques sont limitées dans leur identification des différentes sous-catégorisations. COMLEX décrit les caractéristiques de 92 sous-catégorisation verbales, 14 adjectivales et 9 nominales. Ces caractéristiques expriment : les différentes catégories grammaticales des unités lexicales (noms, verbes, adjectifs, prépositions, adverbes), les différents compléments que peut admettre chacune de ces unités lexicales, ainsi que les propriétés syntaxiques associées à chaque complément. La figure 2.1 montre des entrées appartenant à COMLEX. Pour représenter les propriétés syntaxiques et lexicales de chaque entrée, les concepteurs de COMLEX ont utilisé différents mots clés ou champs. Pour spécifier la forme d'entrée de l'unité lexicale, COMLEX utilise le champ `:orth`. Par exemple : `:orth "abandon"`. Les champs `verb`, `noun`, `prep`, `adverb` mentionnent les catégories syntaxiques de l'unité lexicale (verbe, nom, préposition, adverbe). Si une unité lexicale possède plusieurs catégories syntaxiques, elle aura plusieurs entrées. C'est le cas par exemple de l'unité lexicale *above* qui est définie deux fois : une fois en tant que préposition (`prep :orth "above"`) et une autre fois en tant que adverbe (`adverb :orth "above"`). Les noms, verbes et adjectifs qui ont une morphologie irrégulière disposent des champs `plural` (pluriel), `past` (prétérit), `past-part` (participe passé), etc. Les entrées qui admettent un ou plusieurs compléments ont le champ `subc` (sous-catégorisation). Par exemple, le verbe *abandon* de la figure 2.1 admet deux compléments :

2. Les projets ACQUILEX 1 et 2 ont été financés par la Commission Européenne. Le but du premier projet était de construire une base de connaissance des lexiques multilingues à partir des dictionnaires qui sont exploitables par machine. Le second projet a étendu ce but en explorant l'utilité des corpus textuels (exploitables par la machine) comme sources d'informations lexicales non codées dans des dictionnaires conventionnels. Le projet ACQUILEX II a pris fin en septembre 1995.

3. Depuis 1965, le Linguistic String Project (LSP) est un projet de recherche dans le domaine du traitement automatique du langage naturel. Depuis 1975, le projet a permis de développer des méthodes d'analyse des sous-langages, appliquées en particulier aux documents médicaux. Les applications développées par ce projet effectuent l'extraction de certaines informations contenues dans les bulletins de sortie des hôpitaux et dans les rapports de visites médicales et établissent un codage automatique de ces informations dans un vocabulaire médical bien défini.

```

(verb      :orth "abandon"   :subc((np-pp :pval ("to")) (np)))
(noun      :orth "abandon"   :features ((countable :pval (with))))
(preposition :orth "above")
(adverb     :orth "above")
(verb      :orth "abstain"   :subc ((intrans)
                                     (pp :pval ("from"))
                                     (p-ing-sc :pval ("from"))))
(verb      :orth "accept"    :subc ((np) (that-s) (np-as-np)))
(noun      :orth "acceptance")
(verb      :orth "build"
           :subc ((np) (np-for-np) (part-np :adval ("up"))))
(verb      :orth "seem"      :subc ((to-inf-rs)))

```

FIGURE 2.1 – Un exemple de plusieurs entrées lexicales appartenant à COMLEX

- Le premier complément a pour valeur `np-pp`. Ce complément est associé au sous-schéma `pval ("to")`.
- Le deuxième complément a pour valeur `np`.

La représentation des compléments se fait suivant un codage basé sur les conventions utilisées dans le lexique verbal de Brandeis [Grimshaw and Jackendoff, 1981] dans lequel, chaque complément est représenté par les codes des constituants qui le composent. Si le complément contient un groupe nominal, ce dernier sera représenté par `np` (noun phrase), s'il contient un groupe verbal, il sera codé par `vp` (verbal phrase), le groupe prépositionnel sera codé par `pp` (prepositional phrase), le groupe adjectival par `adjp` (adjectival phrase) et le groupe adverbial par `advp` (adverbial phrase). Si le complément contient plusieurs constituants, ces derniers seront reliés par des traits d'union. Par exemple, le premier complément (`np-pp`) du verbe *abandon* est composé d'un syntagme nominal suivi d'un groupe prépositionnel (exemple 2.1a) et le deuxième complément (`np`) se compose uniquement d'un syntagme nominal (exemple 2.1b). La même figure montre aussi que l'entrée *accept* possède trois compléments : Le premier (noté `np`) est un groupe nominal (exemple 2.2a). Le deuxième complément `that-s` est composé d'une complétive introduite par *that* (exemple 2.2b) et le troisième complément `np-as-np` est un complément qui contient un groupe nominal suivi par la préposition *as* qui est à son tour suivie par un autre groupe nominal (exemple 2.2c).

(2.1) (a) *I abandoned the car to the junkyard*  
(j'ai abandonné la voiture chez le ferrailleur)

(b) *I abandoned the ship*  
(J'ai abandonné le navire)

(2.2) (a) *I accepted my friend's invitation*

(J'ai accepté l'invitation de mon ami)

(b) *The prime minister accepted that Britain had greatly underestimated the danger*  
(Le Premier ministre a admis que la Grande-Bretagne a fortement sous-estimé le danger)

(c) *the group accepted him as a leader*  
(Le groupe l'a accepté en tant que leader)

D'une façon générale, lorsque les compléments admettent des formes complexes, les champs relatifs à ces dernières sont rajoutés. Ces champs varient suivant les types des compléments. Par exemple, l'entrée *abstain* (Figure 2.1) admet un complément qui se compose d'un groupe prépositionnel (pp). Ce complément admet une préposition (*from*) signalée par le champ *pval* ("from"). Ce dernier signifie que le groupe prépositionnel doit être introduit par la préposition *from* (exemple 2.3a). Cette même entrée est associée à un complément de type gérondif noté *p-ing-sc*. Ce champ signifie que le complément est composé d'une préposition (notée *p*) suivi d'un gérondif (noté *ing*) auquel il est associé une propriété syntaxique notée *sc* pour signifier *subject control*. Ceci veut dire que le sujet, qui est le sujet de la proposition principale, est aussi le sujet du gérondif, qui représente la proposition enchassée (exemple 2.3b). Le champ *p-ing-sc* possède un sous-champ, noté *pval* ("from") qui a la même signification que ci-dessus. L'entrée *build* admet un complément de type *part-np*, ce qui veut dire qu'il se compose d'une particule (*part*) suivie d'un syntagme nominal (*np*). Le sous-champ *adval* ("up") signifie que ce complément admet la particule adverbiale *up* (exemple 2.4).

(2.3) (a) *Dallal has abstained from the vote*  
(Dallal s'est abstenue lors du vote)

(b) *the woman abstained from eating during the X-ray treatment*  
(La femme s'est abstenue de manger pendant son traitement avec les rayons X)

(2.4) – *sediment building up on the ocean floor*  
(des sédiments qui s'accroissent sur le fond de l'océan)

COMLEX traite aussi d'autres caractéristiques particulières à travers l'emploi du champ `:features`. Par exemple, la deuxième entrée *abandon* de la figure 2.1 possède une caractéristique bien spécifique représentée par le sous-champ `countable:pval` ("with"). Ce dernier signifie que lorsque le nom *abandon* est au singulier, il doit être précédé par les déterminants *an* ou *the* (exemples 2.5a , 2.5b et 2.5c), à moins qu'il soit précédé par la préposition *with*, dans ce cas, la présence du déterminant n'est pas requise (exemple 2.5d).

- (2.5) (a) *an abandon of the cyclist*  
 (un abandon du cycliste)  
 (b) *the abandon of the cyclist*  
 (l'abandon du cycliste)  
 (c) *\*abandon of the cyclist*  
 (d) *she danced with abandon*  
 (elle a dansé sans aucune retenue)

Pour décrire les propriétés syntaxiques de chaque complément, les concepteurs de COMLEX ont utilisé des cadres (frames). Quatre propriétés syntaxiques sont étudiées dans chaque cadre, ces propriétés sont représentées par les champs `cs`, `gs`, `features` et `ex` (Cf. Figure 2.2) :

1. le champ `cs` (constituent structure) : il exprime la liste de la structure syntagmatique du complément. Les éléments du champ `cs` sont indexés et leurs indexes sont référencés dans le champ `gs` (grammatical structure).
2. le champ `gs` (grammatical structure) : il représente la structure grammaticale ou la relation fonctionnel entre les constituants. Dans la structure grammaticale des cadres verbaux `vp-frames`, l'index "1" fait toujours référence au sujet grammatical du verbe).
3. le champ `features` (caractéristiques) : il est optionnel et fait référence à différents types de contrôles et d'effacements : subject control (`sc`), subject raising (`sr`), object control (`oc`), etc.
  - subject control (`sc`) : signifie que le sujet de surface est le sujet fonctionnel de la proposition principale et aussi celui de la proposition subordonnée. Dans *John promised to leap over the wall* (John a promis de franchir le mur), *John* est le sujet de la proposition principal et aussi de la proposition subordonnée.

```
(vp-frame to-inf-rs :cs (vp 2 :mood to-infinitive :subject 1)
:features (:raising subject)
:gs (:subject () :comp 2)
:ex "they seemed to wilt.")
```

FIGURE 2.2 – La définition du cadre to-inf-rs dans COMLEX

- subject raising (sr) : signifie que le sujet sémantique de la proposition subordonnée "monte" ou change de position pour devenir le sujet syntaxique de la proposition principale. Le verbe *seem* est un exemple typique d'un subject-raising verb (connu aussi sous les noms de raising to subject verb ou subject-to-subject raising verb). Par exemple, dans la phrase 2.6a, *Bill* est le sujet de la proposition subordonnée. Dans la phrase 2.6b, le sujet *Bill* change de position et devient le sujet du verbe. Les verbes *begin*, *continue*, *appear*, etc. appartiennent aussi à ce type de verbes.

(2.6) (a) *Bill is spearheading the demonstration*

(Bill mène la manifestation)

(b) *Bill seems to be spearheading the demonstration*

(Bill semble mener la manifestation)

- object control (oc) : signifie que l'objet direct dans la proposition principale est l'objet du verbe principal et le sujet de la proposition enchassée. Ainsi, dans la phrase *I advised him to leave the house* (je l'ai conseillé de quitter la maison), *him* est à la fois le COD de la proposition principale *I advised him* et sujet de la proposition subordonnée *to leave the house*.

4. le champ *ex* signale la présence d'un exemple du complément étudié. Par exemple, l'entrée *seem* de la figure 2.1 possède le complément *to-inf-rs* : une infinitive dotée de la propriété syntaxique *subject raising*. La figure 2.2 montre le cadre associé au complément *to-inf-rs*.

Dans la figure 2.2, le champ *vp-frame* signifie que le cadre décrit un groupe verbal (verbal phrase) et que ce dernier est de type *to-inf-rs* (infinitive dotée d'un subject raising). La structure de surface du complément (*cs*) se compose d'un groupe verbal (*vp*) à l'in-

finitif (:mood to-infinitive). Le champ `subject 1` signifie que le sujet fonctionnel (sémantique) est le sujet de cette proposition subordonnée. La propriété syntaxique `raising subject`, qui est déjà présente dans la description du complément (`rs`), est à nouveau mentionnée dans le champ `features (:raising subject)`. Les deux informations ont été incluses dans le cas où la première notation serait plus adéquate que la seconde pour un usage particulier. Dans le champ `gs`, on remarque que `subject` n'est pas indexé, ce qui veut dire que la proposition principale ne possède pas de sujet sémantique.

Lorsque des compléments possèdent des propriétés syntaxiques communes, leurs cadres respectifs sont regroupés dans un `frame-groupe`. Par exemple les compléments : `np-np` (exemple 2.7a), `np-to-np` (exemple 2.7b) et `to-np-np` (exemple 2.7c) possèdent des propriétés syntaxiques communes entre eux car leurs constituants ont les mêmes fonctions syntaxiques et on peut passer d'un schéma à un autre sans altérer le sens général de la phrase. Dans ce cas les cadres associés à ces compléments sont regroupés dans le `cadre-groupe (frame-groupe) np-to-np`. Ce dernier sera défini comme suit : `frame-group np-to-np (*np-np *np-to-np *to-np-np)`. Les compléments précédés par un astérisque signifient qu'ils sont définis dans des cadres syntaxiques individuels.

- (2.7) (a) *I gave my friend a book*  
 (b) *I gave a book to my friend*  
 (c) *I gave to my friend a book*  
 (J'ai donné un livre à mon ami)

**2) COMLEX-PLUS :** COMLEX-PLUS est le résultat de l'enrichissement du contenu de COMLEX dans le but de rajouter certaines informations qui manquaient à ce dernier. COMLEX contient 100 classes verbales qui définissent les propriétés des verbes ainsi que celles de leurs compléments, mais ne contient que très peu de classes nominales et il ne fournit aucune information sur les sous-catégorisations des noms qui admettent des arguments. Pour ajouter toutes ces informations manquantes, les concepteurs de COMLEX-PLUS ont procédé comme suit : ils ont pris le contenu de COMLEX et l'ont enrichi à partir de NOMLEX-PLUS - un lexique de nominalisations dont on verra la description dans la section qui suit - avec de nouvelles entrées nominales. Une fois cette étape terminée, ils ont utilisé des règles très simples pour ajouter, à partir de NOMLEX-PLUS, certaines propriétés sur les sous-catégorisations qui caractérisent les différents noms qui admettent des arguments [Meyers et al., 2004a]. Ces propriétés sont représentées par le type du complément admis par le nom et par les prépositions qui entrent dans la composition de ce complément. L'exemple suivant montre la représentation de l'entrée

abduction dans COMLEX-PLUS. Dans COMLEX, l'entrée nominale (NOUN) *abduction* (la définition de cette entrée dans NOMLEX-PLUS est illustrée par la figure 2.4) n'était associée à aucune information sur le type du complément qu'elle admet, alors que dans COMLEX-PLUS, elle a été enrichie par le type du complément PP (*prepositional phrase*) ainsi que par les différentes prépositions (*of* et *by*) que son groupe prépositionnel admet (*the abduction of the girl by the kidnapper* "L'enlèvement de la fille par le ravisseur")

```
(NOUN      :ORTH "abduction"
      :SUBC( (PP :PVAL( "of" "by" ) ) ) )
```

### 2.2.2 De COMLEX à NOMLEX

NOMLEX et NOMLEX-PLUS sont deux ressources électroniques qui ont été développées dans le cadre du projet Proteus à l'Université de New York. Ce projet a vu le jour dans les années 60 et avait pour objectif de concevoir des systèmes capables de trouver automatiquement certaines informations recherchées par l'utilisateur et de les présenter dans le format demandé par ce dernier. Un des principaux défis de l'équipe qui a travaillé sur ce projet est de doter les ordinateurs de connaissances linguistiques. Les différents types de connaissance que l'équipe a cherché à encoder sont : le vocabulaire, la morphologie, la syntaxe, la sémantique, les propriétés spatio-temporelles, les équivalences de traduction, etc. L'encodage des informations syntaxiques relatives aux nominalisations ainsi qu'à leurs sous-catégorisations a abouti à la création des deux dictionnaires NOMLEX et NOMLEX-PLUS.

**1) NOMLEX (NOMinalization LEXicon) :** NOMLEX [Macleod et al., 1998] est un dictionnaire des nominalisations de l'anglais qui contient 1 025 entrées nominales. Ces entrées ont été sélectionnées parmi les nominalisations les plus fréquentes qui apparaissent dans les corpus suivants : Brown et Wall Street Journal. NOMLEX peut être utilisé dans différentes applications appartenant à des domaines variés : extraction d'informations, traduction automatique, génération de grammaires, correcteurs de grammaires, etc. Ce lexique décrit les compléments admis par les nominalisations et établit les liens qui existent entre les groupes verbaux et les compléments nominaux qui leur correspondent. Ceci est réalisé en reliant : (i) les arguments verbaux principaux (sujet, objet direct et objet indirect) aux valeurs qu'ils peuvent prendre dans les syntagmes nominaux correspondant et (ii) les arguments verbaux obliques aux compléments nominaux obliques. Les valeurs que les arguments nominaux peuvent prendre dans le syntagme nominal sont : déterminant possessif (notés DET-POSS) *his announcement* (son annonce), modificateurs de noms (notés N-N-MOD) *The State Department announcement* (L'annonce du Département d'État), compléments introduits par des prépositions

(notés PP-”prep”) *The announcement of the government* (L’annonce du gouvernement), etc. La syntaxe utilisée pour décrire les compléments des différentes nominalisations est la même que celle qui est utilisée dans COMLEX pour décrire les compléments verbaux (COMLEX est basé sur le codage utilisé dans le lexique verbal de Brandeis). Différents types de nominalisations sont traités par NOMLEX. Il décrit les nominalisations dérivées de verbes, les nominalisations sujets, les nominalisations objets et les nominalisations à base de particules. Les différents types sont présentés ci-dessous :

- VERB-NOM représente les nominalisations issues des verbes. Ce type de nominalisations est le plus fréquent et le plus utilisé : *argue/argument* (se disputer/une dispute), *express/expression* (exprimer/expression), *donate/donation* (faire un don/une donation), etc.). Ex :

(2.8) *They argued about/over how to resolve the problem*  
 (ils se sont disputés sur la façon de résoudre le problème)  
*their argument about/over how to resolve the problem*  
 (leur dispute sur la façon de résoudre le problème)

- VERB-PART représente les nominalisations qui se construisent en concaténant le verbe avec sa particule pour ne former qu’un seul mot : *come back/comeback* (revenir/retour), *count down/countdown* (faire le compte à rebours/compte à rebours), *take off/takeoff* (s’envoler/envol), etc. Ex :

(2.9) *Richard take over this company*  
 (Richard rachète cette compagnie)  
*the Richard takeover of this company*  
 (Le rachat de la compagnie par Richard)

- SUBJECT représente les nominalisations qui ont un rôle sujet en plus de leur rôle de prédicat : *teacher* (professeur), *trader* (négociant), *underwriter* (assureur), etc. Dans l’exemple 2.10b, qui est dérivé de l’exemple 2.10a, le prédicat *teacher* est un sujet en plus de son rôle de prédicat.

(2.10) (a) *[someone] thought Mathematics to Mary*  
 ([Quelqu’un] a enseigné les Maths à Mary)  
 (b) *Mary’s Mathematics teacher*  
 (L’enseignant des Maths de Mary)

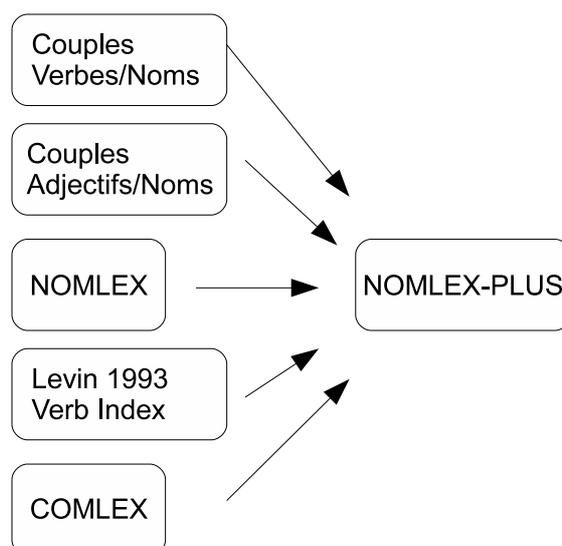


FIGURE 2.3 – Les ressources impliquées dans la création de NOMLEX-PLUS

- OBJECT représente les nominalisations qui ont un rôle d’objet direct en plus de leur rôle de prédicat : *addressee* (destinataire), *appointee* (recrue), *beneficiary* (bénéficiaire), etc. Dans l’exemple 2.11b, le syntagme nominal contient la nominalisation *appointee* qui est aussi l’objet direct du syntagme.

- (2.11) (a) *IBM appointed [someone] as a new Director*  
 (IBM a recruté [quelqu’un] comme nouveau Directeur)
- (b) *IBM’s appointee as a new Director*  
 (La recrue d’IBM comme nouveau Directeur)

2) **NOMLEX-PLUS** : NOMLEX-PLUS [Meyers et al., 2004a] est un lexique créé à partir de plusieurs ressources : COMLEX, NOMLEX, le lexique verbal de Levin (Cf. Figure 2.3). Ce lexique contient 4 900 nominalisations verbales (*destruction, knowledge, believer, etc.*) en incluant les 1 025 nominalisations de NOMLEX, 550 nominalisations dérivées d’adjectifs (*ability, bitterness, etc.*) et 1 600 entrées qui appartiennent à 16 classes contenant des noms ayant des comportements prédicatifs : des noms relationnels (*PRESIDENT of the company*), des noms attributs (*the VOLUME of the sphere*), etc. Dans le but de créer un lexique qui possède une couverture plus large que celle de NOMLEX, les concepteurs de NOMLEX-PLUS ont rajouté des nouvelles nominalisations verbales et adjectivales en utilisant différentes méthodes. Les nom-

```

(NOM
:ORTH          "abduction"
:VERB          "abduct"
:NOM-TYPE      ((VERB-NOM))
:VERB-SUBC
  ((NOM-NP :SUBJECT
    ((DET-POSS)(N-N-MOD)(PP :PVAL ("by"))))
  :OBJECT
    ((DET-POSS)(N-N-MOD)(PP :PVAL ("of")))))

```

FIGURE 2.4 – Un exemple d’une nominalisation appartenant à NOMLEX-PLUS

inalisations verbales ont été principalement déduites à partir de COMLEX, or ce dernier contient des verbes et des noms mais ne fait pas de lien entre les verbes et leurs nominalisations. Pour pouvoir établir cette correspondance, les concepteurs de NOMLEX-PLUS ont utilisé des méthodes de dérivation morphologique dont le but est de comparer les verbes et les noms et d’extraire toutes les paires qui ont des racines communes. Par exemple *destroy* et *destruction* ont en commun la racine *destr*, ou encore *anesthet-ist* et *anesthet-ize*. La méthode qui a été utilisée pour générer des nominalisations adjectivales est la suivante : (i) sont regroupées par paires les nominalisations ainsi que les adjectifs, à partir des différentes ressources qui ont servies à créer NOMLEX-PLUS et qui partagent les mêmes racines, (ii) sont extraits les suffixes qui les caractérisent. (iii) Ces paires de suffixes ont été utilisées pour identifier d’autres couples de nominalisations/adjectifs. Par exemple, les suffixes *-ability/-able* sont extraits à partir du couple *durability/durable* (durabilité/durable) et sont ensuite utilisés pour identifier le couple *availability/available* (disponibilité/disponible). La figure 2.4 montre un exemple qui décrit une nominalisation verbale (*abduction*) appartenant à NOMLEX-PLUS ainsi que les valeurs que peuvent prendre ses arguments nominaux et leurs propriétés syntaxiques.

Dans cet exemple, le champ NOM précise que cette entrée est une nominalisation, le champ ORTH représente la forme de base de cette entrée qui a pour valeur *abduction*, le champ NOM-TYPE précise le type de nominalisation de cet entrée pour laquelle la valeur est VERB-NOM, ce qui signifie que l’entrée est une nominalisation dérivée d’un verbe. Le champ VERB-SUBC spécifie les différents compléments verbaux que le verbe associé à cette nominalisation peut admettre, ainsi que les différentes valeurs que les arguments nominaux peuvent avoir. Dans notre exemple, le verb *abduct* ne peut se construire qu’avec un complément verbal dont la valeur est NOM-NP. Ce complément est formé d’un syntagme nominal qui joue le rôle du

complément d'objet direct. Le champ SUBJECT introduit les valeurs que peut prendre le sujet de la nominalisation dans le schéma nominal. Dans ce cas, le sujet peut être sous les formes suivantes : déterminant-possessif, noté DET-POSS (exemple 2.12a), modifieur de nom, noté N-N-MOD (exemple 2.12b) et complément prépositionnel introduit par la préposition *by*, noté PP : PVAL ( "by" ) (exemple 2.12c). Le champ OBJECT représente les formes sous lesquelles le complément d'objet direct peut apparaître. Il peut être déterminant possessif (exemple 2.12d), modifieur de nom (exemple 2.12e) et complément prépositionnel introduit par la préposition *of* (exemple 2.12f).

- (2.12) (a) *His abduction of Pamela*  
 (Son enlèvement de Pamela)
- (b) *Mike abduction of Pamela*  
 (L'enlèvement de Pamela par Mike)
- (c) *the child abduction by an old men*  
 (L'enlèvement de l'enfant par un vieil homme)
- (d) *James's abduction by two boys*  
 (L'enlèvement de James par deux enfants)
- (e) *the child abduction*  
 (l'enlèvement de l'enfant)
- (f) *the abduction of the woman*  
 (L'enlèvement de la femme)

### 2.2.3 The SPECIALIST Lexicon

The SPECIALIST Lexicon [Browne et al., 2000] (SL) est l'une des trois ressources de l'UMLS qui sont développées par le National Library of Medicine (NLM) et qui font partie du projet Unified Medical Language System. Ce lexique a été développé dans le but de fournir les informations lexicales nécessaires pour les systèmes de traitement du langage naturel spécialisé (SPECIALIST Natural Language Processing System). Il est destiné à être un lexique qui contient des mots de l'anglais général et du domaine Biomédical. Chaque entrée décrit les informations syntaxiques, morphologiques et orthographiques d'un mot donné. Les termes biomédicaux ont été sélectionnés à partir de plusieurs ressources : les résumés appartenant à la Collection de Test de MEDLINE, l'UMLS Metathesaurus et le Dictionnaire Médical Illustré de Dorland. Les termes de l'anglais non spécialisés ont été sélectionnés à partir des ressources de l'American Heritage Word Frequency Book et du Longman's Dictionary of Contemporary

English. Le contenu de ces différentes ressources est composé majoritairement de noms. Pour dériver les verbes et les adjectifs de ces noms, les concepteurs du Specialist Lexicon ont utilisé les heuristiques développées par McCray et al. [1994]. Ce lexique décrit 257 000 entrées (noms, adjectifs, verbes, déterminants ou termes biomédicaux). Il traite 3 788 verbes possédant une ou plusieurs nominalisations ainsi que les schémas de complémentation qui leur correspondent. Nous décrivons ci-dessous les différents types des schémas verbaux répertoriés dans le SL. Pour mieux mettre en évidence les différents arguments impliqués dans les schémas, chaque schéma est associé à un patron syntaxique dont la description est basée sur la notation de Gross [1986], où  $N_0$  représente le sujet,  $V$  le verbe et  $N_1, N_2, N_n$  les différents compléments :

- intrans : correspond au patron syntaxique  $N_0 V$ , c'est-à-dire des verbes sans compléments. Ex : *He disappeared* (il disparaît).
- tran=np : correspond au patron  $N_0 V N_1$ , c'est-à-dire aux verbes transitifs directs qui se construisent avec un complément d'objet direct (COD) et qui acceptent le passif (exemples 2.13a et 2.13b).

(2.13) (a) *little girls prefer dolls*

(les petites filles préfèrent les poupées)

(b) *dolls are preferred by little girls*

(les poupées sont préférées par les petites filles)

- tran=pphr (prep, np) : correspond au patron  $N_0 V \text{ prep } N_1$ , c'est-à-dire aux verbes transitifs dont le complément est introduit par une préposition. Ex : *students gravitate to electronic games* (les étudiants sont attirés par les jeux électroniques).
- ditran=np, pphr (prep, np) : correspond au patron  $N_0 V N_1 \text{ prep } N_2$ , c'est-à-dire aux verbes ditransitifs qui se construisent avec un COD et admettent un second complément introduit par une préposition. Ex : *He gave some money for charity* (il a donné de l'argent pour les bonnes œuvres).
- ditran=pphr (prep, np), pphr (prep, np) : correspond au patron  $N_0 V \text{ prep } N_1 \text{ prep } N_2$ , c'est-à-dire aux verbes ditransitifs qui se construisent avec deux compléments prépositionnels introduits par deux prépositions. Ex : *Leslie emigrated from France to USA* (Leslie a réémigré de la France vers les États-Unis).

Dans l'exemple ci-dessous, nous détaillons la représentation d'un verbe qui appartient au Specialist Lexicon :

```
{base = abate
Entry = E0006436
  Cat = verb
```

```

Variants = reg
intran
tran = np
Nominalization = abatement|noun|E0006437
}

```

Dans cet exemple, l'entrée *abate* admet différents champs et chaque champ possède une ou plusieurs valeurs. Le champ *base* décrit l'entrée de base de l'unité lexicale. Le champ *Entry* représente le numéro d'identification de *abate* dont la valeur est E0006436. Le champ *Cat* représente la catégorie syntaxique qui, ici, est un verbe. Le champ *Variants* représente la flexion du verbe *abate* et la valeur *reg* signale qu'il s'agit d'un verbe régulier, c'est-à-dire, dont le prétérit et le participe passé se construisent avec le suffixe *ed*. Le champ *intran* indique que le verbe est intransitif et donc ne possède pas de complément. Le champ *tran* indique qu'on peut trouver ce verbe dans un schéma transitif avec un groupe nominal dans le rôle du complément d'objet direct. Le dernier champ, *Nominalization*, indique que pour ce verbe, il existe une nominalisation, nommée *abatement*, dont le numéro d'identification a pour valeur E0006437. Ce Lexique contient aussi 3 960 nominalisations. Chaque nominalisation est associée à des compléments de noms introduits par différentes prépositions. Nous donnons ci-dessous, la représentation de la nominalisation *abatement*.

```

{base = abatement
Entry = E0006437
  Cat = noun
  Variants = reg
  Variants = uncount
  Compl = pphr(by,np)
  Compl = pphr(of,np)
  Nominalization_of = abate|verb|E0006436
}

```

La catégorie syntaxique *Cat* indique que cette entrée est un nom (*noun*), dont la morphologie flexionnelle (le champ *Variants*) est décrite avec deux valeurs : (i) *reg* pour spécifier que c'est un nom dénombrable (avec une morphologie flexionnelle régulière) et (ii) *uncount* pour dire que ce nom est aussi un nom non-dénombrable. Les champs *Compl* signifient que *abatement* admet deux compléments de nom : le premier est un complément prépositionnel (*pphr*, pour *prepositional phrase*) pour lequel la préposition *by* introduit un syntagme nominal (*np*), le second est aussi un complément prépositionnel pour lequel la préposition *of* introduit un autre syntagme nominal (*np*). Le dernier champ indique que *abatement* est une nominalisation (*Nominalization\_of*) dérivée du verbe *abate*, qui a le numéro d'identification E0006436. Nous pouvons remarquer dans la description de la nominalisation *abatement* que les compléments (*Compl=pphr(by,np)* et *Compl=pphr(of,np)*) ne donnent

aucune information sur les rôles syntaxiques (sujet, objet, etc.) des syntagmes nominaux qui sont introduits par les prépositions *of* et *by*.

Le SPECIALIST Lexicon nous a servi de base pour l'étude des relations entre les schémas verbaux et les schémas nominaux car il contient des informations très intéressantes à exploiter : (i) il fait le lien entre les verbes et leur nominalisations, (ii) il décrit les schémas verbaux et (iii) il fournit les introducteurs des compléments de noms associés aux groupes verbaux. D'autres part, c'est un lexique bien adapté à la biologie, qui est le domaine sur lequel nous expérimentons des analyses de TAL

### 2.3 Les ressources syntaxico-sémantiques

D'autres travaux se sont intéressés aux propriétés sémantiques des prédicats et ont construit des ressources qui associent propriétés syntaxiques et informations sémantiques dans le but d'améliorer l'analyse du texte. Parmi ces ressources, nous nous intéressons aux trois lexiques suivants : WordNet, VerbNet et FrameNet. WordNet [Miller, 1985, 1995, Fellbaum, 1998] qui est un lexique dans lequel les verbes, les noms, les adjectifs et les adverbes sont regroupés en classes où chaque classe contient les différents synonymes (synsets) de chaque catégorie lexicale. Les classes sont interconnectées au moyen de relations conceptuelles lexicales et sémantiques. VerbNet [Kipper et al., 2000] est un lexique à large couverture dédié aux verbes. Il s'appuie sur les classes de Levin [1993] et en donne une extension. Chaque classe verbale contient différents types d'informations tels que les propriétés syntaxiques (description des arguments) et sémantiques (description des rôles thématiques). Enfin, FrameNet [Johnson et al., 2002] qui est un lexique qui associe les propriétés syntaxiques et sémantiques des prédicats. Il a été créé dans le cadre du projet FrameNet qui a été élaboré à l'Institut de Berkeley (International Computer Science Institute de Berkeley). La caractéristique principale de ce lexique réside dans le fait que sa conception est basée sur la sémantique des cadres (frames) de Fillmore [1982], Fillmore and Atkins [1992]. FrameNet a pour ambition de décrire toutes les possibilités syntaxiques et sémantiques de chaque mot dans chacun de ses sens.

Les corpus sont un autre type de ressources électroniques. Parmi ces ressources : le Penn Proposition Bank [Palmer et al., 2005] (PropBank) et NomBank [Meyers et al., 2004b] concernent les structures prédicatives. Le premier est un corpus annoté avec les groupes verbaux et leurs arguments et le deuxième est un corpus obtenu dans le cadre d'un projet d'annotation des noms (lié au projet PropBank) et qui en donne la structure argumentale.

### 2.3.1 WordNet

WordNet [Miller, 1985, 1995, Fellbaum, 1998] est un lexique sémantique de l'anglais qui a été construit manuellement. Il a été développé en 1985 au sein du Laboratoire des Sciences Cognitives de l'Université de Princeton sous la direction d'un psychologue. Il a été créé dans le but de développer un système basé sur les connaissances acquises à partir des principes psychologiques qui régissent la mémoire lexicale de l'être humain. Les concepteurs ont organisé le lexique d'une façon qui rend compte de la manière dont l'Homme accède naturellement à l'information lexicale. Les linguistes et les chercheurs dans le domaine de l'intelligence artificielle l'ont ensuite utilisé dans le but de produire une ressource qui joue le rôle d'un dictionnaire et d'un thésaurus et qui supporte des applications émanant du domaine de l'analyse automatique du texte ainsi que de l'intelligence artificielle.

WordNet utilise les relations sémantiques d'hyponymie/hyperonymie pour construire un réseau sémantique structuré non pas entre différentes entrées lexicales mais entre les entrées lexicales qui sont reliées entre elles par une relation de synonymie. Le lexique regroupe quatre catégories syntaxiques : noms, verbes, adjectifs et adverbes et chacune d'elle possède sa propre organisation lexicale<sup>4</sup>. Son objectif est de rajouter aux différentes entrées lexicales des propriétés sémantiques. Pour cela, il est organisé d'une façon hiérarchique où les nœuds ne représentent pas des unités lexicales mais des ensembles sémantiques appelés *synsets* (*synonym set*). Un synset est un ensemble qui contient : (i) une liste de synonymes de mots ou de collocations qui représentent le sens du synset, (ii) une définition qui explique le sens du synset et (iii) un ou plusieurs exemples qui impliquent une ou plusieurs unités lexicales. Ce lexique contient 155 327 mots organisés en 117 597 synsets, ainsi qu'un nombre très important de relations sémantiques.

Les mots sont reliés entre eux par des liens lexicaux (antonymie, liens dérivationnels, etc.) et les synsets par le biais de relations sémantiques et conceptuelles. Les relations dépendent du type syntaxique des unités lexicales. Ci-dessous, nous détaillons les relations selon le type syntaxique de l'unité lexicale :

#### (i) Noms

- antonymie : Unités ou termes qui sont voisins et dont les sens sont contraires (*small* vs *big*, *male* vs *female*).
- hyperonymie : Y est hyperonyme de X si chaque X est un type particulier de Y (X

---

4. Cette organisation est justifiée par des tests montrant qu'en règle générale les personnes à qui l'on demande de dire à quoi un mot leur fait penser, répondent par un autre mot qui appartient à la même catégorie syntaxique [Miller, 1990]

plus spécifique et Y plus général). Par exemple l'unité lexicale *red* est hyperonyme des unités lexicales : *scarlet* (écarlate), *vermillion* (vermillon) et *carmine* (carmin).

- hyponymie : Y est hyponyme de X si chaque Y est un type particulier de X. C'est la relation inverse de l'hyperonymie. Par exemple *red* est hyponyme de *color*, *cat* (chat) de *animal* (animal), etc.
- holonymie : relation qui existe entre un terme qui représente la globalité et un terme qui représente une partie, ou un membre de, cette globalité. Par exemple, *tree* (arbre) est un holonyme de *bark* (écorce), de *trunk* (tronc) et de *limb* (branche).
- méronymie : C'est une relation hiérarchique qui existe entre deux concepts ou deux signes linguistiques, dans laquelle le premier est une partie d'un tout que constitue le second. Y est méronyme de X si Y est une partie de X, c'est la relation inverse de l'holonymie. Par exemple : *wheel* (roue) est méronyme de *car* (voiture) et *knee* (genou) de *leg* (jambe).

(ii) Verbes

- hypéronymie : le verbe Y est hypéronyme du verbe X si l'action ou l'activité de X est un type particulier de celle de Y (*to travel* (voyager) et *to move* (bouger))
- troponymie : le verbe Y est troponyme du verbe X si Y est en train de faire X d'une certaine façon ou manière.
  - *to march* (marcher au pas)  $\Rightarrow$  *to walk* (marcher) d'une certaine manière.
  - *to taste* c'est *to eat* d'une certaine manière.
- l'implication (entailment<sup>5</sup>) : le verbe Y est occasionné par le verbe X (ou Y is entailed by X) si l'activité de Y nécessite au préalable l'activité de X. Par exemple, la relation *snore* (ronfler) by *sleep* (dormir) signifie : la véracité de la phrase *a person who is snoring* (une personne qui ronfle) implique que la phrase *this person is sleeping* (cette personne dort) est vraie.
- termes coordonnées : ces verbes partagent un hypéronyme commun. Par exemple les mots *motorbyke* (moto) et *car* (voiture) ont le même hypéronyme *means of conveyance* (moyen de transport).

WordNet présente la particularité d'avoir été utilisé dans un grand nombre d'applications de TAL dont les plus significatives sont :

---

5. Définie dans Crystal [2003] comme : Terme dérivé à partir de la logique formelle et utilisé dans l'étude de la sémantique. Il renvoie à une relation entre une paire de phrases tel que la véracité de la seconde phrase découle nécessairement de celle de la première, par exemple : *I can see a dog*  $\Rightarrow$  *I can see an animal* (Je vois un chien (vraie)  $\Rightarrow$  je vois un animal)

1. des outils de TAL exploitant la sémantique car il est organisé en un réseau sémantique et ne contient que très peu d'informations syntaxiques (structures des arguments verbaux, nominaux, etc.).
2. la création ou l'extension de bases de connaissances. Knight and Luk [1994] ont créé une large ontologie ou base de connaissances destinée à la traduction automatique. Pour cela, ils ont utilisé des méthodes semi-automatiques pour fusionner différentes ressources (WordNet, LDOC, le dictionnaire bilingue Harper-Collins (Espagnol-Anglais), etc.). Green et al. [2001] ont utilisé WordNet pour enrichir automatiquement 4 076 verbes appartenant à la base de données lexicale LVD (Lexical Conceptual Structure Verb Database) avec les sens qui leur correspondent.
3. la génération, ce qui lui permet d'avoir plusieurs relations sémantiques qui sont utiles pour le choix des items lexicaux (lexical choice).

Une représentation partielle de certains noms appartenant à WordNet est illustrée par la figure 2.5, où le synset central représenté par les unités lexicales *father*, *male parent*, *begetter* est en relation d'hyponymie avec quatre synsets *dad*, *father-in-law*, *old man*, *Pater*, etc. et en relation d'hypéronymie avec au moins trois synsets *parent*, *genitor*, *Progenitor*, etc.. Chaque synset contient en plus des différents synonymes, une définition ainsi qu'un exemple.

### 2.3.2 FrameNet

La base de données lexicale FrameNet [Johnson et al., 2002] a été développée pour l'annotation sémantique de corpus de l'anglais, en identifiant les rôles sémantiques liés aux arguments des différents prédicats (verbes, noms, adjectifs, etc.). La conception de FrameNet est basée sur l'utilisation de cadres où les unités lexicales ayant un sens d'utilisation commun sont regroupées dans un seul cadre et celles qui possèdent des sens différents sont mises dans des cadres séparés. Le concept de cadre repose sur les travaux de Fillmore [1982], Fillmore and Atkins [1992]. Ces cadres décrivent la situation conceptuelle qui implique l'une des unités lexicales appartenant au cadre, car selon Fillmore, on peut comprendre le sens d'un mot (unité lexicale) en le reliant à la situation conceptuelle qui lui est sous-jacente. Les concepteurs de FrameNet se sont inspirés de cette notion pour adopter une démarche onomasiologique<sup>6</sup>. Ils se sont intéressés d'abord à la description de la situation (concept) pour construire ensuite la liste des unités lexicales qui expriment un des aspects de cette situation. Une unité lexicale peut être un verbe, un nom, un adjectif, etc.

---

6. Ce terme est défini dans [Dubois et al., 1989] comme "une étude sémantique des dénominations, partant du concept et recherchant les signes linguistiques qui lui correspondent."

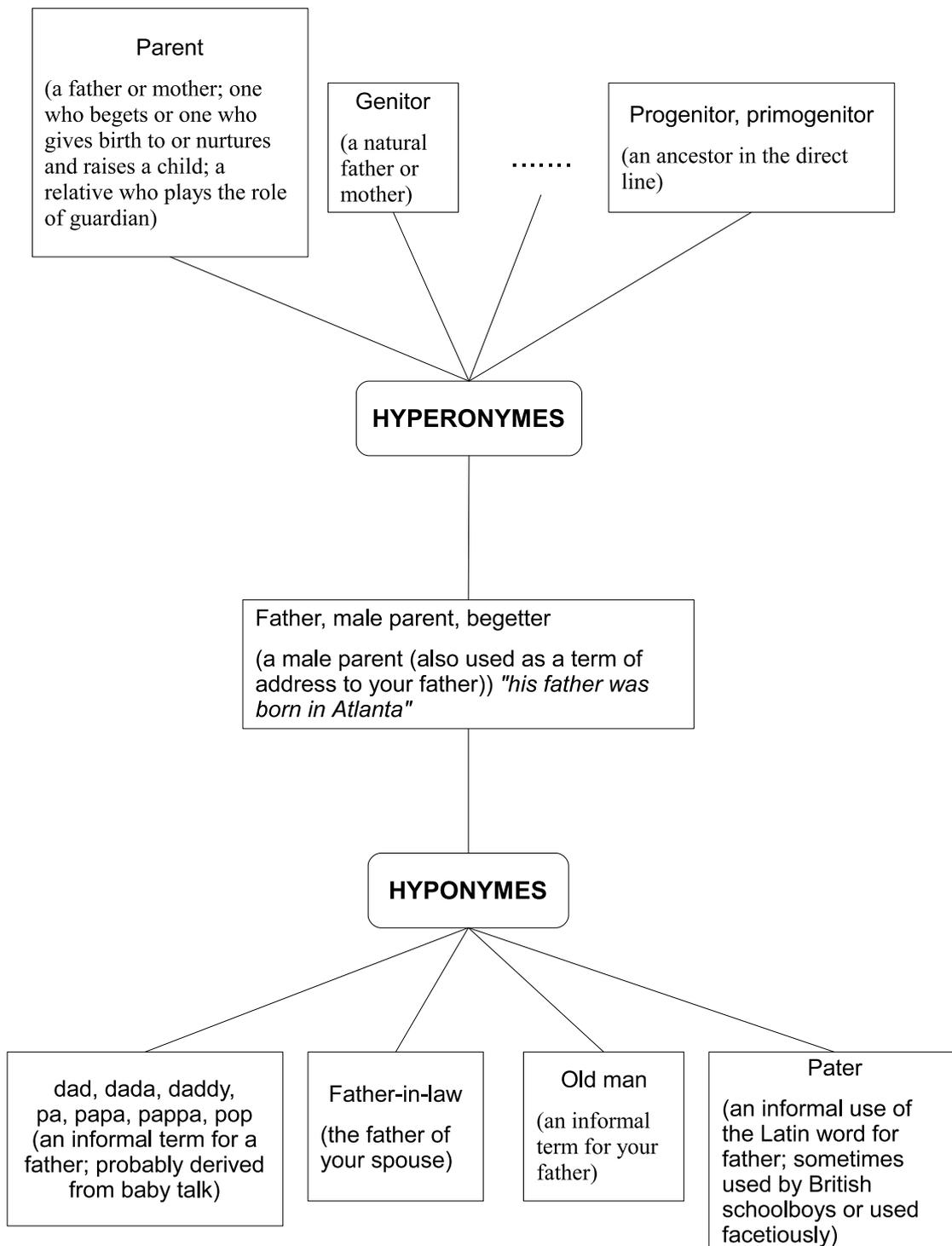


FIGURE 2.5 – Une représentation partielle de la hiérarchie des noms de WordNet

Ainsi, le cadre sémantique *Apply\_heat* (Appliquer une source de chaleur) décrit une situation qui implique plusieurs éléments ou participants. Cette situation est définie comme suit : "A Cook applies heat to Food, where the Temperature\_setting of the heat and Duration of application may be specified. A Heating\_instrument, generally indicated by a locative phrase, may also be expressed. Some cooking methods involve the use of a Medium (e.g. milk or water) by which heat is transferred to the Food. A less semantically prominent Food or Cook is marked Co\_participant"<sup>7</sup>. Dans cette définition, les rôles sémantiques impliqués dans la situation sont soulignés. Ces éléments sont appelés Frame Elements (FEs)<sup>8</sup>. Le cadre contient aussi les différentes unités lexicales (LUs) qui évoquent la situation. Par exemple, *bake* (cuisiner), *blanch* (blêmir), *boil* (bouillir), *broil* (griller), etc. L'exemple 2.14 montre une phrase dont l'unité lexicale *fry* (frir) et les FEs *Cook* (cuisinier), *Food* (nourriture) et *Heating\_instrument* (l'instrument qui fournit ou transmet la chaleur) décrivent une situation qui appartient au cadre sémantique *Apply\_heat* (Appliquer une source de chaleur). Dans cette situation, le FE *Cook* applique une source de chaleur sur une poêle (*Heating\_instrument*) pour frir (*fry*) des frites (*Food*).

(2.14) *[Sally]<sub>Cook</sub> [fried] [*chips*]<sub>Food</sub> [*in a iron skillet*]<sub>Heating\_instrument</sub>*  
 (Sally a frit des pommes de terre dans une poêle de fer)

L'exemple 2.15 illustre une autre situation qui appartient au cadre sémantique *Judgment\_direct\_address*. Cette situation implique l'unité lexicale *scold* (gronder, réprimander, rouspéter) et les différentes unités lexicales qui ont le même sens, tels que *admonish*, *berate*, *chide*, etc. ainsi que les FEs *Communicator* (communicateur) et *Reason* (raison). Ce cadre décrit une situation dans laquelle une personne communique (*communicator*) en donnant son avis ou en émettant une critique sur une autre personne (*Addressee*)<sup>9</sup> pour une raison donnée (*Reason*) ou sur un sujet donné (*Topic*). Cet avis est donné directement à la personne concernée. On remarque dans cet exemple que le communicateur (*Communicator*) a communiqué son désaccord en réprimandant directement les enfants (*Addressee*) sur leur façon de manger les biscuits (*Topic*).

7. Un cuisinier applique une (source de) chaleur à la nourriture où les paramètres de température et le temps d'application peuvent être spécifiés. L'instrument qui permet d'avoir cette source de chaleur est généralement indiqué par un syntagme locatif. Certaines façons de cuisiner nécessitent l'utilisation d'un moyen (par exemple lait ou eau) par lequel la chaleur est transmise à la nourriture. Un élément ou une nourriture dont le sens n'est pas très important pour la compréhension de la phrase est marqué avec le libellé Co\_participant.

8. On peut traduire Frame element par élément de base

9. Qui est le destinataire de la communication

- (2.15) *[the teacher]*<sub>Communicator</sub> SCOLDED *[the children]*<sub>Addressee</sub> *[about their messy cookie eating]*<sub>Topic</sub>  
 (l'instituteur a réprimandé les enfants sur leur façon malpropre de manger les biscuits)

D'une façon générale, un frame est défini dans la sémantique des cadres de Fillmore comme un cadre sémantique ou une structure conceptuelle qui décrit un type particulier de situations ou d'événements ainsi que les différents participants (FEs) impliqués. La figure 2.6 représente le cadre sémantique Revenge (vengeance) où tous les FEs ou participants (Avenger (vengeur), Injured\_Party (partie lésée), Injury (blessure, offense), Offender (coupable), Punishment (punition, châtiment), Degree (le degré), Manner (manière), Purpose (le but)) sont mis en couleur. Chaque couleur représente un FE particulier. Les couleurs sont utilisées aussi pour marquer les participants dans les différents exemples. Le cadre Revenge décrit une situation dans laquelle le vengeur (Avenger) obtient réparation en prenant sa revanche (*retaliate, revenge, avenge, get even, etc.*) pour une offense (Injury) que cette personne ou qu'une tiers personne (Injured\_Party) a subi, de la part du coupable (Offender).

On notera dans ce cadre que les FEs sont distingués en deux classes : Les FEs noyaux (Core FEs) et les FEs non essentiels (Non-Core FEs). Cette classification est faite selon l'importance de l'implication des FEs dans la situation décrite par le cadre sémantique. Les FEs centraux sont ceux dont la présence permet de dire que nous nous trouvons dans une situation bien précise et non pas dans une autre, c'est-à-dire qu'ils participent à l'unicité des situations et par la même occasion à celle des cadres sémantiques. Sans ces FEs, la situation ne serait pas complètement définie. Dans notre exemple (Figure 2.6), Avenger (vengeur), Punishment (punition), Offender (coupable), Injury (offense) et Injured\_party (partie lésée) sont des FEs noyaux, car un acte de vengeance inclut forcément un ou plusieurs éléments de ce type. On ne peut imaginer un acte de vengeance sans qu'il soit précédé par une offense (Injury) ou qu'il ne soit dirigé contre quelqu'un (offender). Les phrases 2.16a, 2.16b et 2.16c montrent des exemples où les différents FEs noyaux sont mis entre crochets.

- (2.16) (a) *[they]*<sub>Avenger</sub> took revenge *[for the deaths of two loyalist prisoners]*<sub>Injury</sub>  
 (ils se sont vengés de la mort de deux prisonniers loyalistes)  
 (b) *[Ryan]*<sub>Avenger</sub> went out to avenge *[them]*<sub>Injured\_Party</sub>  
 (Ryan est sorti pour les venger)  
 (c) *the next day, [the Roman forces]*<sub>Avenger</sub> took revenge *[on their enemies]*<sub>Offender</sub>  
 (le jour suivant, les forces Romaines ont pris leur revanche sur leur ennemi)

Le deuxième type des FEs qu'on peut trouver dans les cadres sémantiques sont les FEs non essentiels. Ce type de FEs peut être présent dans d'autres cadres sémantiques. Par exemple, Time (temps), Place (lieu), Manner (manière), Means (moyen), Degree (degré), etc. sont des FEs qui ne sont pas spécifiques à un cadre donné mais qui appartiennent à plusieurs Frames. La présence seule de ce type de FEs ne permet pas de différencier les cadres (situations) entre eux, mais ils peuvent apporter plus de précision sur une situation donnée. Par exemple, la phrase 2.17 décrit une situation de vengeance, représentée par le prédicat *retaliate* (se venger) d'une famille endeuillée (*bereaved family*). Cet exemple montre que le FE *immediately* (immédiatement) n'est pas nécessaire à la compréhension de la situation, mais apporte des précisions concernant l'instant où s'est déroulée l'action de vengeance. Le cadre sémantique *Revenge* (Figure 2.6) contient la liste de toutes les unités lexicales (UL) - verbales, nominales ou adjectivales - qui décrivent la même situation (revanche), telles que : *avenge*, *get.back(at)*, *payback*, *retaliate*, etc. Et qui impliquent les mêmes FEs (participants sémantiques), qu'ils soient noyaux ou non essentiels : *Avenger* (vengeur), *Punishment* (punition), *Offender* (coupable), *Injury* (offense), *Injured\_party* (partie lésée), *Degree* (degrés), *Manner* (manière), etc.

(2.17) *The bereaved family retaliated [immediately]<sub>Time</sub>*  
 (la famille endeuillée s'est immédiatement vengée)

Chaque cadre sémantique contient des liens vers d'autres cadres qui décrivent les réalisations syntaxiques des différents FEs qui sont associés à chaque unité lexicale, c'est-à-dire, dans quelle position syntaxique un FE associé à une unité lexicale peut se produire et avec quelle préposition (si le FE admet une préposition). Le tableau 2.1 montre les réalisations syntaxiques des FEs *Container*, *Cook*, *Food* et *Heating\_instrument*, lorsqu'ils sont reliés à l'unité lexicale *bake* qui appartient au cadre sémantique *Apply\_heat*. Les réalisations syntaxiques des différents FEs sont les suivantes :

1. Le FE *Container* a été annoté dans deux exemples attestés et possède deux réalisations syntaxiques : la première est mentionnée par le champ *PP[in]* pour signifier que le FE est introduit par la préposition *in* (exemple 2.18a). La deuxième réalisation est signalée par le champ *PP[on]* pour dire qu'il peut aussi être introduit par la préposition *on* (exemple 2.18b).

(2.18) (a) *[Mary]<sub>COOK</sub> baked [cakes]<sub>FOOD</sub> [in beautiful moulds]<sub>CONTAINER</sub>*

## Revenge

### Definition:

This frame concerns the infliction of punishment in return for a wrong suffered. An <b>Avenger</b> performs a <b>Punishment</b> on a <b>Offender</b> as a consequence of an earlier action by the <b>Offender</b> , the <b>Injury</b> . The <b>Avenger</b> inflicting the <b>Punishment</b> need not be the same as the <b>Injured Party</b> who suffered the <b>Injury</b> , but the <b>Avenger</b> does have to share the judgment that the <b>Offender</b> 's action was wrong. The judgment that the <b>Offender</b> had inflicted an <b>Injury</b> is made without regard to the law.	
	(1) <b>They</b> took <b>REVENGE</b> for the deaths of two loyalist prisoners.
	(2) <b>Lachlan</b> went out to <b>AVENGE</b> them.
	(3) The next day, the Roman forces took <b>REVENGE</b> on their enemies.

### FEs:

#### Core:

<b>Avenger [Agt]</b> Semantic Type Sentient	The <b>Avenger</b> exacts revenge from the <b>Offender</b> for the <b>Injury</b> . We want to <b>AVENGE</b> her.
<b>Injured Party [Inrd_prt]</b>	This frame element identifies the constituent that encodes who or what suffered the <b>Injury</b> at the hands of the <b>Offender</b> . Sometimes, an abstract concept such a person's honour or their blood is presented as the element that has suffered the <b>Injury</b> . These also constitute instances of <b>Injured Party</b> . Sam's brothers <b>AVENGED</b> him. We will decide later how to <b>AVENGE</b> the blood of the fallen.
<b>Injury [Injry]</b>	The <b>Injury</b> is the injurious action committed by the <b>Offender</b> against the <b>Injured Party</b> . This Frame Element need not always be realized, although it is conceptually necessary. The team sought <b>REVENGE</b> for their 4-1 defeat last month.
<b>Offender [Off]</b>	The <b>Offender</b> has committed the earlier <b>Injury</b> for which the <b>Avenger</b> seeks revenge. Jo had the affair as a kind of <b>REVENGE</b> against Pat. Marie took terrible <b>REVENGE</b> on Trevor.
<b>Punishment [Pun]</b>	The <b>Avenger</b> carries out a <b>Punishment</b> in order to exact revenge on the <b>Offender</b> . The team took <b>REVENGE</b> with a resounding victory.
<b>Non-Core:</b>	
<b>Degree [Degr]</b> Semantic Type Degree	This FE identifies the <b>Degree</b> to which an event occurs. Marie took <b>terrible</b> <b>REVENGE</b> on Trevor.
<b>Depictive [Depict]</b>	<b>Depictive</b> identifies a phrase describing the actor of an action.
<b>Instrument [Ins]</b> Semantic Type Physical_entity	This FE identifies the <b>Instrument</b> with which the revenge is performed.
<b>Manner [Mann]</b> Semantic Type Manner	This frame element refers to the <b>Manner</b> in which the <b>Avenger</b> exacts their revenge from the <b>Offender</b> . They took <b>brutal</b> <b>REVENGE</b> on the criminal.
<b>Place [Place]</b> Semantic Type Locative_relation	This FE identifies the <b>Place</b> where the revenge occurs.
<b>Purpose [Purp]</b> Semantic Type State_of_affairs	This FE identifies the <b>Purpose</b> for which the revenge is performed.
<b>Result [Result]</b>	This FE identifies the <b>Result</b> of inflicting punishment for a wrong suffered.
<b>Time [Time]</b> Semantic Type Time	This FE identifies the <b>Time</b> when the revenge occurs.

### Lexical Units

*avenge.v, avenger.n, get\_back\_(at).v, get\_even.v, payback.n, retaliate.v, retaliation.n, retribution.n, retributive.a, retributory.a, revenge.n, revenge.v, revengeful.a, revenger.n, sanction.n, vengeance.n, vengeful.a, vindictive.a*

FIGURE 2.6 – Un exemple d'un cadre sémantique de FrameNet

Frame Element	Number Annotated	Realization(s)
Container	(2)	PP[in].Dep (1) PP[on].Dep (1)
Cook	(10)	CNI. -- (10)
Food	(10)	NP.Ext(1) NP.Obj(6) CNI. -- (3)
Heating-Instrument	(8)	INI. -- (7) PP[in].Dep (1)

TABLE 2.1 – Les réalisations syntaxiques de quelques FEs appartenant au cadre sémantique *Apply\_heat*

(Mary a cuit des gâteaux dans de très jolis moules)

(b) *bake [the pancake]<sub>FOOD</sub> [on a preheated baking sheet]<sub>CONTAINER</sub> [for 05-10 min]<sub>DURATION</sub>*

(cuire la crêpe sur une plaque de cuisson préchauffée pendant 5 à 10 minutes).

2. Le FE *Cook* a été annoté dans dix exemples. [CNI<sup>10</sup>. --] est la seule réalisation syntaxique rattachée à ce FE, ce qui signifie que le rôle sémantique *Cook* est effacé dans les dix exemples qui ont été annotés. Dans ce cas, le FE *Cook* peut avoir différentes fonctions syntaxiques :

- sujet dans les phrases impératives (exemple 2.19). On remarque dans cet exemple que le FE *COOK* est effacé.

(2.19) *bake [the tart]<sub>FOOD</sub> [on a preheated baking sheet]<sub>CONTAINER</sub> [at 350°F (180°C) gas mark 4]<sub>TEMPERATURE</sub> [for 40-45 min until the filling is creamily set]<sub>DURATION</sub>*

(cuire la tarte sur une plaque de cuisson préchauffée à 180°C pendant 40 à 45 minutes jusqu'à ce que la garniture soit crémeuse)

- agent dans les phrases passives (exemple 2.20). Cet exemple se caractérise également par l'effacement du FE *COOK*.

10. Constructional Null Instantiation (CNI) est le cas où un constituant qui appartient à une construction est omis. Dans ce cas, le constituant est précisé et mis entre crochets. Le CNI est utilisé dans les cas suivants : omission du sujet d'une phrase impérative, omission de l'agent d'une phrase passive, omission de l'objet d'une phrase impérative qui contient des instructions

(2.20) – [*The mix*]<sub>FOOD</sub> is baked [*for 20 minutes*]<sub>DURATION</sub> [*in moulds*]<sub>CONTAINER</sub> and served with a vegetable cream sauce  
 (le mélange est cuit au four pendant 20 minutes dans des moules et servi avec une sauce à la crème de légumes)

3. Le troisième FE (*FOOD*) est annoté dans dix exemples. Il est marqué avec le champ [NP . Ext ] dans un seul exemple. Dans ce cas, il représente un groupe nominal qui joue le rôle d'un argument externe (exemple 2.21a). Il est aussi marqué par le champ [NP . Obj ] dans six exemples, ce qui veut dire que le FE *FOOD* joue le rôle du complément d'objet direct dans tous ces exemples (exemple 2.21b). Enfin, il est marqué par le champ [CNI . -- ] dans trois exemples pour signifier que dans ce cas, *FOOD* est effacé (exemple 2.21c).

(2.21) (a) [*the bread*]<sub>Food</sub> has been baked [*in an oven*]<sub>Instrument.heating</sub>  
 (le pain a été cuit dans un four)  
 (b) [*Henry*]<sub>COOK</sub> baked [*the potatoes*]<sub>FOOD</sub>, then opened them lengthways  
 (Henry a cuit les pommes de terre puis les a ouverts en longueur)  
 (c) *bake* [*at 180C*]<sub>TEMPERATURE</sub> [*for 30 minutes*]<sub>DURATION</sub>  
 (cuire pendant 30 minutes à la température de 180°).

4. Le dernier FE *Heating\_Instrument* est annoté dans huit exemples. Il est marqué par le champ PP[ in ] .Dep<sup>11</sup> dans un seul exemple pour signifier qu'il est introduit par la préposition *in* (exemple 2.22a). Dans les sept autres exemples, il possède la propriété [INI<sup>12</sup> . - ]. Cette propriété signifie que le FE *Heating\_Instrument* est volontairement omis (exemple 2.22b).

(2.22) (a) *We baked the cake* [*in the oven*]<sub>Heating\_Instrument</sub>  
 (Nous avons cuit le gâteau dans le four)  
 (b) *bake* [*the soufflé*]<sub>FOOD</sub> [*for 12 minutes*]<sub>DURATION</sub>  
 (cuis le soufflé pendant 12 minutes)

11. Le sous-champ *Dep* représente une fonction grammaticale générale attribuée aux adverbes, groupes prépositionnels, groupes verbaux ou propositions qui se produisent après leur principal prédicat (verbes, adjectifs ou noms) dans une phrase déclarative simple. Cette fonction grammaticale fait référence à tout ce qui est argument ou complément adverbial

12. Indefinite Null Instantiation (INI) est le cas où des verbes transitifs comme *eat* (manger), *sew* (coudre), *bake* (cuire), etc. perdent leur compléments d'objet direct et se comportent comme des verbes intransitifs (*Molly rarely eats alone*, Molly mange rarement seule).

FrameNet contient plus de 10 000 unités lexicales, dont 6 100 sont complètement annotées<sup>13</sup> dans plus de 825 cadres sémantiques ou conceptuels. Il contient aussi environ 700 FEs. L'ensemble des frames sont conçus de façon hiérarchique et leur construction s'est faite manuellement en se basant sur l'intuition des personnes dont l'anglais est la langue maternelle. En utilisant des méthodes automatiques pour générer les cadres sémantiques, Green et al. [2004], Green and Dorr [2004] ont développé SemFrame qui est un lexique très similaire à FrameNet et dont les objectifs sont les mêmes.

### 2.3.3 VerbNet

VerbNet [Kipper et al., 2000] est un lexique des verbes de l'anglais qui possède une très large couverture. Il se caractérise par la présence de liens qui le relient avec d'autres outils et ressources tels que : WordNet, XTAG<sup>14</sup> [Doran et al., 1994], une plate-forme d'analyse reposant sur le formalisme TAG [Joshi et al., 1975], et FrameNet. VerbNet est compatible avec WordNet tout en proposant plus d'informations syntaxiques et sémantiques. Il s'appuie sur les classes de Levin en les enrichissant avec d'autres sous-classes. Chaque sens verbal est associé à une classe, si un verbe possède plusieurs sens, il référencera plusieurs classes. Par exemple, le verbe *run* (courir) dans *they ran out of the house* (ils sont sortis de la maison en courant) appartient à la classe *Manner of Motion* (manière de mouvement), alors que *run* dans *the railway line runs through a valley* (la voie ferrée parcourt la vallée) appartient à la classe *Meander* (serpenter, faire des méandres). VerbNet est relié à WordNet dans le but d'avoir un accès à des définitions plus précises des différents verbes. Ce lien est établi en reliant chaque classe verbale (une classe représente un sens d'un verbe donné) au(x) synset(s) appartenant à WordNet qui ont le sens le plus proche du sens de cette classe. Chaque classe verbale possède : (i) un ensemble de descriptions syntaxiques (cadres syntaxiques) et (ii) des restrictions sémantiques. Les descriptions syntaxiques, qui ont été capturées en LTAG (Lexicalized Tree Adjoining Grammar) [Joshi, 1987, Schabes, 1990], décrivent les réalisations syntaxiques (surface realization) de la structure argumentale des différentes constructions : transitives, intransitives, prépositionnelles, résultatives, etc., un certain ensemble de constructions décrivant les variations ou les alternances de diathèse (voix), ainsi que la nature syntaxique des constituants rattachés aux rôles thématiques et dans certains cas, les prépositions admises.

---

13. Annoter les unités lexicales revient à spécifier tous les FEs qui leur sont liés, à préciser leurs types (Core ou non-Core) et à donner des exemples de phrases montrant les différents FEs ainsi que leurs caractéristiques syntaxiques (les positions syntaxiques qu'ils peuvent avoir dans une phrase donnée)

14. XTAG est un projet qui a pour but de développer une grammaire de large couverture en utilisant le formalisme TAG (Tree Adjoining Grammar). XTAG est aussi un système pour le développement de TAGs et se compose d'un analyseur syntaxique, d'une interface X-windows pour le développement de la grammaire et d'un analyseur morphologique

Les restrictions sémantiques imposent des contraintes aux membres de chaque classe sur leur comportement sémantique. Ces restrictions sont représentées par :

1. des rôles thématiques <sup>15</sup> pour les structures argumentales prédicatives que les membres de cette classe autorisent. Par exemple, le verbe *lend* (prêter) est considéré comme un verbe trivalent car il peut se construire avec trois arguments.
2. des restrictions de sélection (ou règles sélectionnelles <sup>16</sup>) qui définissent le type de chaque argument. Ces restrictions sont de différents types :
  - des restrictions utilisées pour contrôler le type des rôles thématiques tolérés par les arguments (*animate* (animé), *human* (humain), *concrete* (concret), etc.) Par exemple, le rôle thématique *instrument* (instrument) associé à l'unité lexicale *kick* (donner un coup de pied) doit être de type *foot* (pied), ou encore, que le complément d'objet direct (*Patient*) du verbe *kill* (tuer) soit de type animé.
  - des restrictions utilisées pour contrôler la direction d'un verbe de mouvement (*directed motion*). Dans l'exemple *Tom rolled the ball down the hill* (Tom a fait rouler le ballon jusqu'au bas de la colline), *roll* est un verbe de mouvement et sa direction est vers le bas.

Les cadres syntaxiques sont associés à des informations sémantiques, exprimées sous la forme d'une conjonction de plusieurs prédicats sémantiques de type primitifs 'motion', 'contact' ou 'cause'. Chaque prédicat est associé à un événement E décomposé en une structure tripartite inspirée des travaux de Moens and Steedman [1988]. Cette structure représente trois phases appartenant à l'événement E :

- *processus préparatoire* (*preparatory process*) : représenté par le prédicat sémantique *during*(E) et qui correspond à toute la phase qui précède le changement d'état.
- *the culmination* : représenté par *end*(E) et qui correspond à l'instant (point ponctuel) où s'opère le changement d'état.
- *l'état résultat* ou *consequent state* : correspond au nouvel état qui survient juste après le point de culmination. Cet état est représenté par *result*(E).

---

15. Les rôles thématiques (ou  $\theta$ -roles) sont le moyen formel pour représenter la structure syntaxique des arguments (leurs nombres et leurs types syntaxiques) exigés par le verbe.

16. Les règles sélectionnelles spécifient les restrictions qui s'appliquent sur les combinaisons possibles des différentes unités lexicales dans un contexte grammatical donné.

Une fonction est introduite pour chaque prédicat spécifiant si ce dernier est vrai ou non durant une étape donnée. Par exemple, la phrase *Stephen climbed to the top* (Stephen a grimpé jusqu'au sommet) représente une situation qui a duré dans le temps (l'action d'escalader) et qui s'est terminée par un point de culmination représenté par *the top* (le sommet) où peut s'opérer un changement d'état. Un autre exemple qui illustre la notion de décomposition des événements est représenté par le verbe *break* (casser). Dans l'exemple *Maya broke the glass* (Maya a cassé le verre), on peut faire la distinction entre l'état de l'objet en question avant la fin de l'action (*during(E)*), c'est-à-dire l'état et la forme du verre avant de le casser, et le nouvel état qui résulte après (*end(E)*, *result(E)*), c'est-à-dire les débris de verre résultant. Ainsi, la structure tripartite des événements permet d'exprimer la sémantique des classes qui représentent les verbes d'état (*climb (escalader)*, *break (casser)*, etc.). La figure 2.7 détaille les propriétés syntaxiques et les contraintes sémantiques associées aux membres de la classe *Hit* où :

- le champ MEMBERS représente la liste de tous les membres (prédicats) appartenant à la classe *hit* (frapper, heurter, percuter), c'est-à-dire les prédicats qui véhiculent un sens identique *hit*, *kick* (donner un coup de pied), *slap* (donner une gifle), etc.
- le champ THEMATIC ROLES liste tous les rôles thématiques qui peuvent être associés aux schémas syntaxiques des différents prédicats. Par exemple, les membres de la classe *Hit* peuvent avoir les rôles thématiques Agent, Patient et Instrument.
- le champ SELECT RESTRICTIONS représente les restrictions sélectionnelles appliquées sur les rôles thématiques. Dans le cas de notre exemple, la propriété Agent[+animate] signifie que l'agent doit être "animé", la propriété Patient[+concrete] signifie que l'objet doit être "concret" et la propriété Instrument[+concrete, -animate] signifie est que "l'instrument" est un objet concret mais non animé.
- le champ FRAMES and PREDICATES représente les cadres syntaxiques qui contiennent les différents schémas syntaxiques ainsi que les prédicats sémantiques qui leur sont associés et qui sont tolérés par cette classe.

Le dernier champ (FRAMES and PREDICATES) est représenté par trois colonnes, chacune d'elles référence une propriété particulière : la première colonne représente les différentes constructions permises par la classe : Resultative, Conative<sup>17</sup>, Basic Transitive (transitif avec COD), Transitive with Instrument (transitif indirect avec un instrument introduit par *with*), against/on alternation (une construction où on peut alterner

17. Le *conatif* est un type de formation verbale propre à exprimer l'effort ; ainsi, en français l'imparfait peut être un conatif [Dubois et al., 1991]

HIT class

⟨⟨MEMBERS ⟩⟩		[[ <i>hit</i> , 1), ⟨ <i>kick</i> , 1), ⟨ <i>slap</i> , 1), ⟨ <i>tap</i> , 1), ...]
⟨⟨THEMATIC ROLES ⟩⟩		Agent(A), Patient(P), Instrument(I)
⟨⟨SELECT RESTRICTIONS ⟩⟩		Agent[+animate], Patient[+concrete], Instrument[+concrete,-animate]
⟨⟨FRAMES and PREDICATES ⟩⟩		
Basic Transitive	A V P	manner(during(E),directedmotion,A) ∧ manner(end(E),forceful,A) ∧ contact(end(E),A,P)
Transitive with Instrument	A V P with I	manner(during(E),directedmotion,I) ∧ manner(end(E),forceful,I) ∧ contact(end(E),I,P)
Together reciprocal	A V P[+plural] together	manner(during(E),directedmotion,P <sub>i</sub> ) ∧ manner(during(E),directedmotion,P <sub>j</sub> ) ∧ manner(end(E),forceful,P <sub>i</sub> ) ∧ manner(end(E),forceful,P <sub>j</sub> ) ∧ contact(end(E),P <sub>i</sub> ,P <sub>j</sub> )
Resultative	A V P Adj	manner(during(E),directedmotion,A) ∧ manner(end(E),forceful,A) ∧ contact(end(E),A,P) ∧ Pred(result(E),P)
Resultative	A V P Adj with I	manner(during(E),directedmotion,I) ∧ manner(end(E),forceful,I) ∧ contact(end(E),I,P) ∧ Pred(result(E),P)
Resultative	A V P PP	manner(during(E),directedmotion,A) ∧ manner(end(E),forceful,A) ∧ contact(end(E),A,P) ∧ Pred(result(E),P)
Body-part object or reflexive object	A V I[+body-part/+refl] against/on P	manner(during(E),directedmotion,I) ∧ manner(end(E),forceful,I) ∧ contact(end(E),I,P)
Conative	A V at P	manner(during(E),directedmotion,A)
Conative	A V at P with I	manner(during(E),directedmotion,I)
With/against alternation	A V I against/on P	manner(during(E),directedmotion,I) ∧ manner(end(E),forceful,I) ∧ contact(end(E),I,P)
Body-part object or reflexive object	A V I[+body-part/+refl]	manner(during(E),directedmotion,I) ∧ manner(end(E),forceful,I) ∧ contact(end(E),I,?)

FIGURE 2.7 – Un exemple d’une classe verbale de VerbNet

les prépositions *against* et *on*), *body-part object*, *Together reciprocal*, etc. La deuxième colonne représente les schémas syntaxiques qui correspondent aux constructions mentionnées dans la première colonne, certains schémas syntaxiques sont détaillés ci-dessous :

- le schéma A V P : correspond à la construction *Basic Transitive* où le A est l'agent et le P est le thème ou patient (exemple 2.23).

(2.23) *the children kicked the ball*  
(l'enfant a donné un coup de pied au ballon)

- le schéma A V P with I : correspond à la construction *Transitive with Instrument*. Une construction qui admet la préposition *with* comme introducteur d'un instrument I (exemple 2.24).

(2.24) *Ryan hit his sister with a stick*  
(Ryan a frappé sa sœur avec un bâton)

- le schéma A V P Adj : correspond à une construction *résultative* associée à un adjectif (exemple 2.25).

(2.25) *Mike kicked the door open*  
(Mike a ouvert la porte d'un coup de pied)

- le schéma A V P PP : correspond à une construction *résultative* associée à un groupe prépositionnel PP (exemple 2.26).

(2.26) *Mike kicked the ball into the net*  
(Mike a envoyé le ballon au fond de la cage (avec le pied)).

- le schéma A V at P : correspond à une construction *conative* qui nécessite la préposition *at* comme introducteur du patient P (exemple 2.27).

(2.27) *the robber hit at the policeman face's*  
(Le voleur a attaqué le policier au visage)

- le schéma A V I against/on P : correspond à la construction *against/on* alternation dans laquelle le patient P est introduit soit par la préposition *against* soit par *on*.

(2.28) *the little girl hit the stick against/on the table*  
(la petite fille a frappé le bâton contre la table)

- le schéma  $A\ V\ I[+body-part/+refl]\ against/on\ P$  : correspond à une construction dans laquelle l'instrument  $I$  peut être de type *body-part object* lorsque le COD est une partie d'un corps (*tête, coude, etc.*) ou de type *reflexive object*<sup>18</sup> et le patient  $P$  est introduit par les prépositions *against* ou *on* (exemple 2.29).

(2.29) *he was tapping his hands on the desk*  
(il frappait ses mains sur le bureau)

Cette même colonne contient des restrictions appliquées sur les schémas syntaxiques et les constructions qui leur correspondent. Par exemple, la classe *Hit* ne peut permettre la construction conative que si la préposition *at* est présente ( $A\ V\ at\ P$  et  $A\ V\ at\ P\ with\ I$ ). Une autre construction qui nécessite des restrictions est la construction *Together reciprocal alternation* ( $A\ V\ P[+plural]\ together$ ), qui est une construction qui admet un patient  $P$  suivi par *together*. Cette restriction exige que le patient soit au pluriel *John hits the sticks together* (John frappe les bâtons l'un contre l'autre). La troisième colonne représente un ensemble de restrictions sémantiques exprimées sous la forme d'une conjonction de plusieurs prédicats primitifs. Le rôle des prédicats est de restreindre le comportement sémantique des différents rôles thématiques. Par exemple, les rôles thématiques liés au schéma syntaxique *Basic Transitive* ( $A\ V\ P$ ) sont associés aux prédicats : *manner* et *contact*. Le prédicat *manner* est utilisé une première fois pour signifier que durant l'événement  $E$  ( $during(E)$ ), l'agent  $A$  est dans un mouvement orienté (*directed motion*). Ce même prédicat est utilisé une seconde fois pour signifier qu'à la fin de l'événement  $E$  ( $end(E)$ ), l'agent  $A$  va avoir un comportement agressif (*forceful*). Le prédicat *contact* représente un contact établi à la fin de l'événement  $E$  entre l'agent  $A$  et le patient  $P$ . La conjonction entre ces différents prédicats signifie que durant l'événement  $E$ , l'agent  $A$  est en mouvement et qu'à la fin de cet événement, l'agent  $A$  établit un contact ou une relation avec le patient  $P$  d'une manière forte ou agressive.

D'après la figure 2.7, on remarque que les prédicats sémantiques associés aux constructions conatives ne traitent pas la fin d'un événement ( $end(E)$ ). Ceci s'explique par le fait que ce genre de constructions ne permet d'atteindre aucun but (pas de résultat). Le prédicat sémantique  $Pred(result(E), P)$  associé aux constructions résultatives signifie que le résultat de l'événement  $E$  est représenté par le patient  $P$ . Le schéma syntaxique

---

18. Lorsque le sujet et l'objet ont le même référent *I dressed* (je me suis habillé), *the man shaved* (l'homme s'est rasé)

A V I[+body-part/+refl] qui est relié à la construction *Body-part object* ou *reflexive object* possède deux prédicats sémantiques *manner* et *contact*. La conjonction de ces derniers signifie que durant l'événement *E*, l'instrument *I* est en mouvement (*manner(during(E), directedmotion, I)*) et qu'à la fin de cet événement, l'instrument *I* est impliqué dans un contact brutal *manner(end(E), forceful, I)*, mais qu'on ignore avec quoi ou contre quoi l'instrument est entré en contact *contact(end(E), I, ?)*. Par exemple dans la phrase *Caroline hit her head* (Caroline a heurté sa tête), on ne sait pas contre quoi l'agent *Caroline* a frappé sa tête (qui joue le rôle de l'instrument).

### 2.3.4 PropBank

Le Penn Proposition Bank [Palmer et al., 2005], appelé aussi PropBank, est un corpus annoté en anglais et utilisé pour la reconnaissance des groupes verbaux, il a été développé dans le but de rajouter des informations sémantiques au Penn TreeBank [Marcus, 1994, Marcus et al., 1993]. Son développement a été fait en 2000, à l'Université du Colorado et avec l'appui de l'ACE Program<sup>19</sup>. Il a impliqué plusieurs équipes de recherche : BBN<sup>20</sup>, MITRE<sup>21</sup>, les universités de New York et de Pennsylvanie. Ce corpus concerne essentiellement les structures argumentales des prédicats verbaux. L'enrichissement du Penn TreeBank est réalisé avec une couche d'annotations sémantiques (rôles sémantiques) en complément de la structure syntaxique qui est déjà présente. Il ignore le rôle des adjectifs et des prédicats nominaux.

Le point de départ de PropBank a été l'utilisation d'un corpus contenant un million de mots. Ce corpus est le Penn Treebank II Wall Street Journal. L'objectif principal de l'annotation réalisée par PropBank est de trouver le rôle adéquat lorsqu'il existe plusieurs formes syntaxiques impliquant le même verbe. Une analyse syntaxique simple des exemples 2.30a et 2.30b montre que *window* est le complément d'objet direct du premier exemple et le sujet du deuxième mais cette analyse ne montre pas si cet argument possède le même rôle sémantique ou non. Les annotateurs humains de PropBank ont utilisé les représentations des structures propositionnelles des phrases appartenant au Penn TreeBank et les ont transformées en structures prédicats-arguments (Predicate argument structure, PAS). Pour cela, ils ont associé à chaque entrée lexicale verbale un cadre qui inclut ses éventuels compléments ou arguments ainsi que les rôles sémantiques associés à chaque argument. Par exemple, Arg0 pour référencer l'agent et Arg1 pour le thème. Les

19. L'objectif de ce programme est de développer une technologie d'extraction automatique de contenus dans le but de l'utiliser en TAL

20. BBN est une entreprise qui a été fondée en 1948 par des chercheurs du MIT. Son objectif initial était de donner des conseils dans le domaine de l'acoustique.

21. MITRE a été fondée en 1958 comme une association sous la direction de C.W. Halligan. Elle se spécialise dans l'ingénierie des systèmes, la technologie de l'information, la conception de l'exploitation (Systèmes de transport) et la modernisation des entreprises

cadres contiennent aussi des informations concernant les relations qui lient les arguments entre eux. Par exemple, la structure argumentale liée au verbe *keep* dans la phrase *I kept the ice cream in the freezer* (j'ai gardé la glace dans le congélateur) est : [arg0 [keep arg1 arg2]] où il existe une relation entre arg1 et arg2. Il est à noter aussi que les annotations de PropBank ne font pas de distinction entre les arguments obligatoires et les arguments optionnels.

(2.30) (a) [ARG0 John] broke [ARG1 the window]

(John a cassé la fenêtre)

(b) [ARG1 The window] broke

(la fenêtre a cassé)

### 2.3.5 NomBank

NomBank [Meyers et al., 2004b] est un projet d'annotation développé à l'Université de New York. Il est en lien avec le projet PropBank de l'Université du Colorado. NomBank fait partie d'un vaste effort qui consiste à ajouter différentes couches d'annotation au Penn Treebank II dans le but de fournir aux chercheurs travaillant dans le domaine du TAL de meilleurs outils pour l'analyse automatique du texte. L'imprécision et la fragilité des analyseurs à couches (composés de plusieurs transducteurs mis en cascade) des décennies passées ont accéléré la volonté de remplacer ces derniers par des analyseurs basés sur les treebanks qui possèdent des performances meilleures mais dont l'analyse syntaxique est plus superficielle (shallow parsing<sup>22</sup> [Meyers et al., 2004b]). NomBank a pour objectif d'annoter les structures argumentales des noms appartenant au corpus Penn TreeBank, comme PropBank l'a fait déjà avec les arguments verbaux. L'objectif initial du projet NomBank avait pour ambition de fournir la structure argumentale de 5 000 noms communs. Pour cela, il s'est appuyé essentiellement sur le projet NOMLEX car la plupart des noms qui admettent des arguments sont des nominalisations ou des noms qui ont les mêmes propriétés que les nominalisations. Il traite aussi d'autres phénomènes comme : les constructions à base de verbes supports (exemple 2.31a), les constructions qui contiennent des copules (exemple 2.31b) ou encore des incises qui sont introduites par des prépositions (exemples 2.31c).

(2.31) (a) Mark made a decision

(Mark a pris une décision)

(b) her decision is to immigrate

(sa décision est d'émigrer)

---

22. Les analyseurs à base de treebank dépendent des treebanks sur lesquels ils reposent. Comme la nature de ces derniers est plutôt adaptée à l'analyse syntaxique superficielle, le résultat va forcément être de la même nature.

- (c) *with hostilities continuing, the state of emergency was extended indefinitely in February 1993 at the president's request*  
 (avec les troubles qui continuent, l'état d'urgence a été prolongé indéfiniment en février 1993 par ordre du président)

En plus des nominalisations dérivées de verbes (*decision* (décision), *helper* (quelqu'un qui aide), *nominee* (personne nommée), etc.), NomBank traite les nominalisations issues d'adjectifs (*incompetence* (incompétence), *ability* (capacité), *wisdom* (jugement), etc.) ainsi que d'autres types de noms qui admettent des arguments. La figure 2.8 montre des exemples de phrases annotées suivant le codage adopté par les concepteurs de NomBank. Le champ [NOM] de l'exemple 2.32a

- (2.32) (a) *His gift of a car to William* [NOM]  
 REL = *gift*, ARG0 = *his*, ARG1 = *a car*, ARG2 = *to William*  
 (son cadeau (une voiture) à William)
- (b) *his promise to make the travel more comfortable* [NOM]  
 REL = *promise*, ARG0 = *his*, ARG2-PRD = *to make the travel more comfortable*  
 (sa promesse de rendre le voyage plus confortable)
- (c) *their only hope of keeping customers from defecting to other competitors*  
 REL = *hope*, ARG0 = *their*, ARG1-PRD = *of keeping customers from defecting to other competitors*  
 (son seul espoir d'empêcher les consommateurs d'aller à la concurrence)
- (d) *Mindy's mother* [DEFREL]  
 REL = *mother*, ARG0 = *mother*, ARG1 = *Mindy*  
 (la mère de Mindy)
- (e) *A possible U.S. troop increase in Irak*  
 REL = *increase*, ARG1 = *U.S. troop*, ARGM-LOC = *in Irak*, ARGM-ADV = *possible*  
 (une possible augmentation des troupes américaines en Irak)
- (f) *Another small encouragement to the Federal Reserve to lower interest rates in coming weeks* [NOM]  
 REL = *encouragement*, ARG1 = *to the Federal Reserve*, ARG2-PRD = *to lower interest rates*, ARGM-TMP = *in coming weeks*  
 (d'autres petits encouragements pour la réserve fédérale pour abaisser les taux d'intérêt dans les semaines à venir)
- (g) *TOTAL made an agreement* [NOM/SUPPORT]  
 SUPPORT = *made*, REL = *agreement*, ARG0 = *TOTAL*  
 (Total a passé un accord)
- (h) *The six-mile trip to my airport hotel*  
 REL = *trip*, ARG3 = *to my airport hotel*, ARGM-EXT = *six-mile*  
 (les six miles qui me séparent de mon hôtel)

FIGURE 2.8 – Des exemples annotés appartenant à NomBank

signifie que le prédicat *gift* (signalé par le mot clé REL) est une nominalisation dérivée à partir d'un verbe. Cette nominalisation admet comme sujet (ARG0) le déterminant *his*, le complément d'objet direct *a car* est signalé par le champ ARG1 et le complément d'objet indirect *to William* est introduit par le champ ARG2<sup>23</sup>. La nominalisation *promise* de l'exemple 2.32b est aussi une nominalisation verbale (NOM), le déterminant *his* est le sujet (ARG0) et le deuxième complément (ARG2) *to make the travel more comfortable* est accompagné de la propriété -PRD. Cette propriété est utilisée dans cet exemple et dans l'exemple 2.32f pour signaler une infinitive. Le champ -PRD sert aussi à signaler les gérondives, c'est le cas de l'exemple 2.32c. Lorsque le verbe d'un exemple admet un COD (exemple 2.32f), le champ PRD se construit avec ARG2 et dans le cas contraire (exemple 2.32b), le champ PRD se construit avec ARG1. L'exemple 2.32d est étiqueté avec DEFREL pour signaler que le nom *mother* est un nom relationnel. L'exemple 2.32e met en évidence un autre type d'argument, en l'occurrence, ARGM-LOC. Le champ ARGM marque la présence d'un complément circonstanciel. Quand le complément circonstanciel est locatif *in Irak*, le champ ARGM est associé avec le champ LOC. Quand le complément circonstanciel est un argument modifieur, adverbial ou adjectival (*possible* dans notre exemple), le champ ARGM-ADV est utilisé. On remarque dans l'exemple 2.32f que le complément *in coming weeks* est marqué avec le champ ARGM-TMP. Dans ce cas, ce complément est un complément circonstanciel temporel. L'exemple 2.32g est marqué par l'étiquette NOM/SUPPORT pour signaler que la nominalisation *agreement* est introduite par le verbe support *made*. Le champ ARGM-EXT de l'exemple 2.32h signifie que *the six-miles* est un complément circonstanciel de mesure. En général, ce champ est présent lorsque la phrase contient un complément circonstanciel de mesure, de degré, d'intervalle, etc.

Toutes ces propriétés combinées aux propriétés de PropBank mettent à disposition de l'utilisateur des données annotées afin de reconnaître les régularités lexicales et syntaxiques associées aux structures des phrases ainsi qu'à celles des syntagmes nominaux. Ceci permet de n'utiliser qu'un seul schéma pour déduire les variations de diathèse. Par exemple, à partir de la phrase *IBM appointed John* (IBM a embauché John) où l'argument objet (ARG1) est *John* et où le sujet (ARG0) est représenté par *IBM*, un système d'analyse basé sur NomBank et PropBank peut déduire que IBM a embauché John à partir de : *John was appointed by IBM* (John a été embauché par IBM), *IBM's appointment of John*, *The appointment of John by IBM* (L'embauche de John par IBM) et *John is the current IBM appointee* (John est l'actuel recrue d'IBM).

---

23. ARG2 représente soit le complément d'objet indirect, soit le complément prépositionnel

## 2.4 Conclusion

Nous avons vu dans ce chapitre qu'il existe une grande variété de ressources électroniques. Le nombre de ces ressources est en constante augmentation car elles sont exploitées par de nombreuses applications du domaine du TAL (traduction automatique, recherche documentaire, résumés de textes, extraction d'informations, questions réponses, analyseurs syntaxiques et sémantiques, etc.). Pour avoir des résultats précis et optimaux, les ressources électroniques doivent être bien structurées et dans des formats permettant aux différentes applications de les exploiter au mieux. Pour bien montrer qu'il existe différents types de ressources, nous avons réparti ces dernières en deux grandes classes : la première classe contient les ressources qui étudient les propriétés lexico-syntaxiques des différentes unités lexicales et la deuxième classe contient les ressources qui s'intéressent aux propriétés lexico-sémantiques et syntaxico-sémantiques.

COMLEX et COMLEX-PLUS sont deux ressources lexico-syntaxiques concernant principalement les propriétés syntaxiques des verbes et celles des compléments verbaux. COMLEX étudie certaines propriétés syntaxiques liées aux autres catégories lexicales (noms, adjectifs, ad- verbes), mais d'une façon limitée. Nous avons vu que les concepteurs de COMLEX ont voulu combler un manque. Ce dernier ne traitait pas suffisamment les propriétés syntaxiques des nominalisations et surtout celles des compléments nominaux qui leur sont associés. Pour cette raison ils ont développé COMLEX-PLUS. Ce dernier se caractérise principalement par la présence d'informations supplémentaires sur les compléments prépositionnels admis par certains noms ainsi que leurs introducteurs (prépositions), mais cette information reste insuffisante.

NOMLEX et NOMLEX-PLUS ont été développés par la même équipe pour s'intéresser exclusivement aux propriétés syntaxiques des nominalisations et de leurs compléments nominaux. NOMLEX possède 1 025 nominalisations de différents types : nominalisations verbales, nominalisations sujets, nominalisations objets et des nominalisations à particules, ce qui est insuffisant pour faire de NOMLEX un lexique à large couverture car il existe d'autres types de noms qui admettent des arguments. NOMLEX-PLUS possède une couverture plus large que NOMLEX. En plus des 1 025 nominalisations déjà contenues dans NOMLEX, NOMLEX-PLUS a été enrichi avec 3 800 autres nominalisations de différents types : nominalisations issues d'adjectifs, noms relationnels, noms attributs, etc.

La dernière ressource lexico-syntaxique est The SPECIALIST Lexicon. La couverture de ce lexique est plus large que celles des lexiques vus jusqu'à présent : il contient 257 000 entrées (noms, adjectifs, verbes, nominalisations verbales, nominalisations adjectivales, etc.). Il se distingue des autres lexiques par le fait qu'il contient des termes (noms, nominalisations, verbes) qui appartiennent au domaine biomédical, ce qui représente un intérêt particulier pour notre

travail et ce qui a motivé notre choix pour ce lexique.

La deuxième classe des ressources électroniques que nous avons vus regroupe deux types de ressources : les ressources lexico-sémantiques et les ressources syntaxico-sémantiques. Le seul lexique qui appartient au premier type est WordNet. Ce dernier est un lexique pour l'anglais qui contient à peu près 155 000 mots reliés entre eux par le biais de relations lexicales et sémantiques (synonymie, hypéronymie, hyponymie, etc.). Les ressources syntaxico-sémantiques sont beaucoup plus nombreuses. FrameNet est une ressource développée pour l'annotation sémantique des corpus. Elle est utilisée pour identifier les rôles sémantiques liés aux différents prédicats (verbes, noms, adjectifs, etc.). La conception de FrameNet est fondée sur les cadres de Fillmore. En tant que ressource syntaxico-sémantique, VerbNet qui est considéré comme un des lexiques les plus importants étudie les propriétés syntaxiques et sémantiques des différents prédicats verbaux. Il est organisé en un ensemble de classes verbales fondées sur la classification de Levin où chaque classe est associée à un sens particulier d'un verbe. VerbNet n'étudie que les propriétés syntaxiques des verbes et leurs schémas de complémentation.

En plus des lexiques électroniques déjà cités, il existe aussi des corpus électroniques. PropBank est un corpus annoté utilisé principalement pour la reconnaissance des groupes verbaux et l'annotation des structures argumentales des prédicats verbaux. Il a été développé dans le but de rajouter des informations sémantiques (rôles sémantiques) au Penn TreeBank. PropBank, contrairement à NomBank, ne s'intéresse pas aux adjectifs, ni aux nominalisations ni aux structures argumentales des différents noms. NomBank est un autre type de corpus qui s'intéresse aux argument nominaux. Il est en relation avec le projet PropBank et son but est d'annoter les structures argumentales des noms qui appartiennent au Penn TreeBank. NomBank annoté les nominalisations verbales, adjectivales ainsi que tous les noms qui admettent des arguments. Le tableau 2.9 est un récapitulatif de l'ensemble des ressources électroniques étudiées dans ce chapitre. Toutes ces ressources sont classées selon leurs types : lexiques lexico-syntaxique, lexiques syntaxico-sémantique et corpus. Le tableau décrit : les différentes lexiques qui ont été utilisés dans la création des ressources, les types des données qui sont traitées par chacune de ces ressources, dans quel but elles ont été créées ainsi que les méthodes qui ont été utilisées pour les créer, etc. Nous avons également mis dans ce tableau les caractéristiques du lexique qui a été créé à l'aide de la plate-forme PredicateDB (Cf. Chapitre 4).

Dans ce chapitre, nous avons donné une liste non exhaustive des différentes ressources électroniques qui décrivent les propriétés lexicales, syntaxiques et sémantiques des différents prédicats verbaux et nominaux. Nous avons également vu que certaines ressources s'intéressent aux prédicats nominaux et à leurs structures argumentales. Parmi ces ressources, nous avons choisi d'utiliser le Specialist Lexicon car il possède une large couverture et traite différents

Ressources	Ressources lexico-syntaxiques					Ressources syntaxico-sémantiques			Corpus		Ressource syntaxique
Nom	COMLEX	COMLEX-PLUS	NOMLEX	NOMLEX-PLUS	Specialist Lexicon	WordNet	VerbNet	FrameNet	PropBank	NomBank	PredicateDB
Organisme	New York Univ.	New York Univ.	New York Univ.	New York Univ.	National Library of Medecine (NLM)	Princeton Univ.	Colorado Univ.	Berkeley Univ.	Colorado Univ. + Pennsylvany Univ.	New York Univ.	LIF - CNRS - Aix-Marseille Univ.
Langue et domaine	anglais général	anglais général	anglais général	anglais général	anglais général + biomédical	anglais général	anglais général	anglais général	anglais général	anglais général	anglais général + biomédical
Taille Nb entrées	38 000	non connue	1 025	7050	257 000	155 327	274 classes	+ 11 600	1 million	202 965	3 981
Ressources utilisées	Oxford Advanced Learner's Dictionary	COMLEX + NOMLEX-PLUS	Brown Corpus + Wall Street Journal	COMLEX + NOMLEX + Lexique Verbal de Levin	UMLS metathesaurus + dictionnaire de Dorland + American Heritage Word Frequency + LDOC	non connue	classes verbales de Levin	non connue	Penn TreeBank	le Penn TreeBank	Specialist Lexicon
Type des données	verbes + adjectifs + noms	Verbes + adjectifs + noms + compléments des noms qui en admettent	nominalisations verbales	nominalisations verbales et adjectivales + noms avec arguments	Noms + verbes + adjectifs + nominalisations verbales et adjectivales	Verbes + noms + adjectifs + adverbes	verbes	Verbes + noms + adjectifs	verbes	noms qui possèdent des arguments	nominalisations verbales
Objectifs	étude des propriétés syntaxiques des différents mots + compléments	identique à COMLEX + étude des schémas syntaxiques des SA des nominalisations verbales	étude des schémas verbaux + SA des schémas nominaux associés	étude des schémas verbaux + SA des schémas nominaux associés	étude des propriétés lexicales et syntaxiques	étude des liens sémantiques des différentes entrées	réalisation syntaxique et sémantique des SA	description de la valence de chaque mot selon son sens	annotation des structures argumentales verbales	description des structures argumentales des noms	description syntaxique des structures argumentales des nominalisations
Méthode d'acquisition	manuelle	semi-automatique	manuelle	semi-automatique	non connue	manuelle	semi-automatique	manuelle	manuelle	manuelle + automatique	semi-automatique

FIGURE 2.9 – Les différentes ressources électroniques étudiées

prédicats qui appartiennent au domaine biomédical. Par contre, il fournit une information incomplète et partielle concernant les schémas nominaux associés aux nominalisations. Dans la suite de notre travail (*Cf.* 4), nous complétons l'information qui manque en rajoutant les marqueurs de rôles que peuvent jouer les différentes prépositions fournies par le SL et qui sont présentes dans la description des différents prédicats nominaux et verbaux.



## **Chapitre 3**

# **Structures Prédicatives : leur acquisition et leur capture en corpus**

### 3.1 Introduction

La connaissance des propriétés syntaxiques et sémantiques qui caractérisent les structures argumentales permet de capturer des informations fondamentales sur les relations qui existent entre les participants et les événements. Par exemple, des travaux en analyse syntaxique et génération [Bangalore and Joshi, 1999, Stede, 1998], en traduction automatique [Dorr, 1997], en recherche d'informations [Klavans and Kan, 1998] et en extraction d'informations [Riloff and Schmelzenbach, 1998] s'appuient sur ces données. Capturer les structures argumentales et déduire les relations qui existent entre les prédicats et leurs arguments n'est pas une tâche aisée du fait que le langage naturel dispose de plusieurs réalisations syntaxiques différentes pour exprimer un sens donné. Par exemple, la phrase 3.1a peut être réécrite sous plusieurs formes qui possèdent des structures de surface différentes mais qui préservent le sens général. Elle peut être réécrite : (i) sous une forme passive (exemple 3.1b) où on remplace le verbe *govern* par sa forme passive (auxiliaire *be* + participe passé du même verbe) ou (ii) sous des formes nominales (exemples 3.1c à 3.1e) où le prédicat verbal a été remplacé par une nominalisation qui est un nom dérivé du même prédicat verbal. Il est possible de ramener ces différentes formes de surface à la structure argumentale<sup>1</sup> *govern(the valve, the fuel intake)*. L'alternance du datif est un autre phénomène linguistique qui engendre différentes réalisations syntaxiques. Il ne concerne pas tous les verbes, mais seulement certaines classes verbales telles que *give* (donner), *sell* (vendre), *lend* (prêter) [Levin, 1993](page 45). L'alternance du datif en anglais se caractérise par la possibilité d'inverser l'ordre du thème et du destinataire tout en effaçant la préposition qui est habituellement associée à ce dernier. Par exemple, la phrase 3.1f admet un destinataire (*Henry*) qui est précédé de la préposition *to*. Cette phrase est strictement équivalente à la phrase de l'exemple 3.1g dans laquelle la préposition *to* a été supprimée et le destinataire *Henry* est positionné devant le complément d'objet direct (*ball*). Les phrases de la première forme (avec la préposition) sont nommées datives prépositionnelles alors que celles de la deuxième forme sont appelées datives à double objet.

- (3.1) (a) *the valve governs the fuel intake*  
(la soupape contrôle l'admission du carburant)  
(b) *the fuel intake was governed by the valve*  
(l'admission du carburant est contrôlée par la soupape)  
(c) *the government of the fuel intake by the valve*  
(d) *the fuel intake's government by the valve*  
(e) *the valve's government of the fuel intake*

---

1. La structure argumentale relie l'action ou le prédicat à ses participants

- (le contrôle de l'admission du carburant par la soupape)
- (f) *Zidane gave the ball to Henry*
- (g) *Zidane gave Henry the ball*  
(Zidane a donné le ballon à Henry)

Dans ce chapitre, nous nous intéressons aux structures prédicatives. Dans les Section 2 et 3, nous montrons que la mise en évidence des prédicats ainsi que les liens qui les relient à leurs arguments est très utile en TAL (extraction d'informations, recherche d'informations, etc.). Dans la section 4, on va montrer les différents types d'approches (morphologiques et morpho-sémantiques) qui ont été utilisées dans la capture des structures prédicatives.

### 3.2 Les structures argumentales prédicatives

Après avoir utilisé des analyseurs basés sur des banques d'arbres (treebanks) qui traitent les structures de surface (le Penn Treebank I [Marcus et al., 1993]), les principaux objectifs des chercheurs sont de développer des analyseurs plus efficaces basés sur des banques d'arbres qui implémentent des schémas d'annotation syntaxiques capables de traiter les structures argumentales prédicatives (SAP). Le Penn Treebank II est une banque d'arbres qui implémente ce genre de schéma d'annotation [Marcus et al., 1994]. L'étude des SAP permet de capturer les régularisations syntaxiques, c'est-à-dire qu'elle permet de trouver une représentation commune qui représente différentes formes syntaxiques (forme active, forme passive, etc.) exprimant la même relation sémantique. D'une façon générale, la régularisation syntaxique signifie l'expression d'un ensemble de formes non-canoniques par une forme canonique. Ceci a pour avantage de réduire le nombre des schémas nécessaires pour capturer les relations sémantiques utilisées par différentes applications du TAL [Meyers et al., 2002]. La capture de SAP permet de mettre en évidence de nombreuses informations sur les relations qui existent entre les prédicats et leurs arguments. Parmi ces informations, on peut trouver : (1) des étiquettes de fonction, qui ont pour rôle de classer sémantiquement les différents constituants ou de leur affecter des rôles grammaticaux. (2) des arcs libellés qui représentent les dépendances ou les rôles grammaticaux et qui montrent la façon avec laquelle les constituants sont reliés entre eux, etc. D'autres travaux ont cherché à s'appuyer sur les SAP dans différents domaines : Marcus et al. [1994] les ont employées pour enrichir la forme des annotations utilisées dans le Penn Treebank, Surdeanu et al. [2003] ont proposé une méthode d'extraction d'informations qui exploite les SAP. Selon ces auteurs, cette méthode permet d'obtenir des résultats de meilleurs qualités. Yakushiji et al. [2004] ont proposé une méthode qui extrait les prédicats verbaux à partir des SAP qui résultent d'une analyse grammaticale complète et les utilisent pour faire de l'extraction d'informations

dans le domaine biomédical, ou encore, [Meyers et al., 2002] qui ont affiné les annotations du Penn Treebank II avec des SAP encodées avec le codage utilisé dans le GLARF<sup>2</sup> [Meyers et al., 2001b,a]. Ce dernier est un cadre de représentation grammaticale et logique des arguments destiné à produire des annotations détaillées. Par exemple, il permet d'assigner à chaque constituant un rôle grammatical en le reliant à d'autres constituants de la phrase.

### 3.3 Les prédicats

De nombreuses applications en TAL reposent sur l'acquisition automatique des connaissances lexicales et syntaxiques [Boguraev and Pustejovsky, 1996]. Ces connaissances concernent différents types de prédicats : (i) des prédicats verbaux tels que *inscribe* (inscrire, graver), *acidify* (acidifier) et *absolve* (absoudre), (ii) des nominalisations verbales telles que *inscription* (inscription), *acidification* (acidification), *absolution* (absolution), (iii) des nominalisations adjectivales telles que *abeyance* (suspension) , *abhorrence* (horreur), *ability* (capacité) ou encore des prédicats nominaux qui ne possèdent aucun lien dérivationnel. Les prédicats sont des fonctions générales qui spécifient les relations qui peuvent exister entre certains ensembles de concepts [Norman and Rumelhart, 1975]. Ils ont la capacité de relier les différents arguments (sujet/complément(s)) dans une phrase. La mise en évidence des prédicats ainsi que les liens qui les relient à leurs arguments a permis de développer de nombreuses applications en TAL (Extraction d'Informations, Recherche d'Informations, Résumés Automatique, etc.). Par exemple, Thomas et al. [2000], Sekimizu et al. [1998] se sont intéressés à l'extraction des relations qui existent entre différentes entités à partir d'un texte biologique. Bien que les prédicats nominaux (nominalisations ou autres) soient nombreux dans les textes écrits et spécialement dans les textes scientifiques, tels que les textes biologiques auxquels nous nous sommes intéressés, il n'y a pas eu beaucoup de travaux en linguistique informatique qui se sont intéressés aux propriétés de ce type de prédicats. La plupart des travaux sur les structures prédictives ont porté sur les propriétés syntaxiques et sémantiques des prédicats verbaux.

#### 3.3.1 L'acquisition Automatique de Cadres de Sous-Catégorisation

Différents travaux se sont basés sur les structures argumentales prédictives pour acquérir automatiquement des cadres de sous-catégorisation :

Brent [1993] a développé une méthode automatique qui extrait les propriétés syntaxiques des verbes à partir d'un lexique et les regroupe dans des cadres syntaxiques. Pour pouvoir extraire ces propriétés, l'auteur s'est basé sur certains indices syntaxiques et lexicaux (déterminants,

---

2. Grammatical and Logical Argument Representation Framework

auxiliaires, modaux, prépositions, conjonctions de coordination, etc.) qui appartiennent au schéma du verbe étudié. Par exemple, le verbe *want* (vouloir) peut admettre une infinitive et le verbe *hope* (espérer) admet une complétive, mais le contraire n'est pas permis (exemples 1 à 4). Ce travail a pour objectif de créer des cadres syntaxiques regroupant les verbes qui admettent les mêmes types de compléments verbaux (infinitifs, groupes nominaux, relatifs, etc.) et qui décrivent aussi leurs propriétés syntaxiques.

- (3.2) (a) *Jean wants Fred to be punctual*  
(Jean veut que Fred soit ponctuel)  
(b) *Jean hopes that Fred is punctual*  
(Jean espère que Fred est ponctuel)  
(c) \**Jean wants that Fred is relax*  
(\*Jean veut que Fred est relaxe)  
(d) \**Jean hopes Fred to be relax*  
(\*Jean espère que Fred soit relaxe)

Brent [1991, 1993], Manning [1993] et Ushioda et al. [1993] ont conçu des systèmes d'acquisition simples, capables de reconnaître un nombre réduit de classes verbales de sous-catégorisations (16 classes au maximum). Par la suite, Briscoe and Carroll [1997] ont développé une technique et implémenté un système pour construire automatiquement un dictionnaire de cadres de sous-catégorisation à partir d'un corpus textuel. Ce système est capable de distinguer 160 cadres de sous-catégorisations verbales. Les classes reconnues par ces systèmes peuvent contenir différentes informations : les différents contrôles admis par les arguments (contrôle sur le sujet, sur l'objet, etc.), les différentes alternations comme le mouvement de la particule, etc.

Dorr [1997] a développé une technique similaire qui permet de construire automatiquement des lexiques utilisés dans l'enseignement assisté de langues étrangères ainsi que dans les systèmes de traduction automatique multilingue. Ces dictionnaires sont basés sur les structures conceptuelles et lexicales (LCS<sup>3</sup>) qui sont un langage indépendant de représentation. Selon l'auteur, l'utilisation des LCS a pour objectif de montrer que les verbes synonymes possèdent les mêmes caractéristiques syntaxiques.

---

3. Lexical Conceptual Structure

**Les Structures Conceptuelles et Lexicales :** Une LCS est une représentation hiérarchique abstraite avec des propriétés indépendantes du langage [Traum and Habash, 2000]. Selon Dorr [1994], une LCS est une version modifiée de la représentation proposée par Jackendoff [1983, 1990]. Les LCS sont fréquemment utilisées dans les systèmes de traduction automatique. Par exemple, elles ont été utilisées comme un langage pivot dans les systèmes UNITRAN (UNIversal TRANslator) [Dorr, 1987] - qui est un système syntaxique de traduction automatique multilingue bidirectionnel (anglais, allemand et espagnol) - et MILT (Military Language Tutor) [Dorr et al., 1997] - qui est aussi un système de traduction automatique spécialisé dans le domaine militaire. Dans le système MILT, les LCS ont été associées à des composants définitionnels qui incluent des liens bilingues (issus d'EuroWordNet<sup>4</sup>) entre les mots des langages source et destination.

Une LCS peut être représentée comme un graphe orienté avec une racine. Chaque nœud fait référence à plusieurs informations : un type, une primitive et un champ. Un nœud peut être de différents types : un événement (Event), un chemin (Path), une manière (Manner), un endroit (Location), une chose (Thing), etc. À chaque type est associé un ensemble de primitives structurelles - tels que CAUSE, GO, BE, TO, etc. - qui font référence aux mêmes concepts exprimés par les types. Par exemple la primitive GO exprime le concept de partir (*to go*) et elle est de type événement (Event). La primitive TO est utilisée lorsqu'on veut exprimer un déplacement et elle est de type chemin (Path). Une LCS permet de capturer les propriétés sémantiques d'un item lexical grâce à la combinaison de (i) sa structure sémantique qui est spécifiée par la forme du graphe et qui est héritée à partir de sa classe verbale décrite dans la classification verbale de Levin [1993], et (ii) son contenu sémantique qui est représenté par le sens du verbe lui-même. La LCS de la phrase *John went happily to school* [Dorr, 1994] est représentée comme suit :

```
[Event GOLoc
  ([Thing JOHN],
   [Path TOLoc ([Position ATLoc ([Thing JOHN], [Location SCHOOL])])])
  [Manner HAPPILY]])
```

Le prédicat de cet exemple est la primitive GO qui est de type Event et qui est associé à un lieu (Loc). La primitive admet un sujet logique (JOHN), de type Thing (Chose) et un argument logique, représenté par la primitive TO qui est de type Path (Chemin). Cet argument logique est une LCS imbriquée qui lie le sujet JOHN avec une destination (SCHOOL) de type Location (Lieu) en utilisant la primitive AT de type Position. HAPPILY (heureux) est

4. EuroWordNet une base de données multilingue associée à un ensemble de WordNets pour plusieurs langues (l'italien, l'espagnol, l'allemand, le français, le tchèque et l'estonien)

le modifieur de la primitive GO, il est de type *Manner* (manière). En d'autres termes, cette représentation signifie que GO admet un sujet (JOHN) qui va suivre un chemin (Path) pour aller à (TO) l'école (SCHOOL) qui est de type *Location* (lieu). Ce lieu est une *Position* représentée par AT, dans laquelle le sujet va se trouver. La manière avec laquelle le sujet s'est rendu à l'école est la joie (happily).

Lapata [1999] et McCarthy and Korhonen [1998] se sont intéressées à l'acquisition automatique des sous-catégorisations verbales dans le cas de variation de diathèses. Lapata [1999] s'est intéressée à l'identification des variations de diathèses dans les différents corpus de données. Son travail a consisté à extraire automatiquement les verbes qui peuvent admettre différentes diathèses (datif, bénéfactif, etc.) à partir du British National Corpus [Burnard, 1995] (BNC) en utilisant des méthodes d'analyse partielles et des informations taxinomiques. Comme le montrent les exemples 3.3a et 3.3b, le datif se caractérise par une alternance entre le schéma verbal  $\langle \text{NP V NP}_1 \text{ to NP}_2 \rangle$ , qui admet un complément prépositionnel, et le schéma verbal  $\langle \text{NP V NP}_2 \text{ NP}_1 \rangle$  qui admet deux objets (un complément d'objet indirect et un complément d'objet direct). Le bénéfactif possède une structure similaire au datif, la différence porte sur l'utilisation de la préposition *for* à la place de *to* (exemples 3.3c et 3.3d). McCarthy and Korhonen [1998] ont développé une méthode d'acquisition automatique des variations de diathèses d'un ensemble de verbes issus du British National Corpus.

- (3.3) (a) *John offers a rose to his wife*  
(b) *John offers his wife a rose*  
(John offre une rose à sa femme)  
(c) *leave a tip for the waitress*  
(d) *leave the waitress a tip*  
(laisse un pourboire au serveur)

### 3.3.2 L'acquisition automatique des contraintes de sélection

Différents travaux se sont intéressés à l'acquisition automatique des contraintes de sélection : Resnik [1996] a présenté un modèle formel qui aide à comprendre les relations qui lient les prédicats à leurs arguments et qui permet d'acquérir automatiquement les contraintes de sélection. Ces dernières sont définies comme étant les limitations sur l'applicabilité des prédicats aux arguments, c'est-à-dire, les limitations qui contraignent les prédicats à accepter un type particulier d'arguments et non un autre type. Riloff and Schmelzenbach [1998] ont développé un

algorithme basé sur les corpus pour l'acquisition empirique des cadres casuels conceptuels munis de contraintes de sélection à partir d'un texte non-annoté. En se basant sur des schémas d'extraction et d'un lexique sémantique spécifique à un domaine particulier, l'algorithme est capable d'apprendre les propriétés sémantiques associées à chaque schéma d'extraction et de fusionner ceux qui sont syntaxiquement compatibles dans le but de produire des cadres casuels multi-champs munis de contraintes de sélection. Les cadres casuels contiennent des champs pour les rôles thématiques qui sont associés à chaque événement. Par exemple, les cadres casuels associés aux activités commerciales contiennent des champs qui représentent les agents (les entreprises ou les personnes qui fusionnent ou qui acquièrent d'autres entreprises) et les objets (les entreprises qui ont été acquises ou les produits qui ont été développés).

### 3.3.3 L'acquisition d'informations sémantiques sur les aspects verbaux

L'aspect verbal est une catégorie grammaticale utilisée dans la description des verbes qui ont un mode et un temps et référençant principalement la façon dont la grammaire marque la durée ou le type de l'activité temporelle indiquée par le verbe [Crystal, 2003]. Par exemple, le contraste qui existe entre l'accompli (perfectif ou parfait) et le non-accompli (ou imperfectif). Différents travaux se sont intéressés à l'acquisition d'informations sémantiques sur les aspects verbaux :

Klavans and Chodorow [1992] ont présenté une méthodologie qui permet d'obtenir des informations sémantiques sur les aspects des verbes. Ils ont proposé une représentation de l'aspect du verbe qui associe une valeur pour les différents types d'événements où cette valeur reflète l'usage typique du verbe. Les auteurs se sont uniquement intéressés aux verbes statifs tels que : *know* (savoir), *resemble* (ressembler), *be* (être), *love* (aimer) et aux verbes non statifs - appelés aussi verbes duratifs - tels que : *run* (courir), *walk* (marcher), *give* (donner), *open* (ouvrir). L'objectif des auteurs était de donner un degré de stativité/non-stativité aux différents verbes. Leurs résultats ont été obtenus de deux manières : (i) en utilisant le corpus de Brown [Francis and Kucera, 1982] qui est un corpus étiqueté et (ii) en appliquant l'analyseur syntaxique English Slot Grammar<sup>5</sup> (ESG) [McCord, 1980, 1990], qui annote les constituants des phrases avec leur nature et leur rôle thématique, sur le corpus du Reader's Digest<sup>6</sup>

Siegel [1999] a développé un système complet de classification verbale basé sur la classification aspectuelle. Ce système utilise des indicateurs linguistiques représentés par des marqueurs lexico-syntaxiques qui sont linguistiquement liés aux différents aspects verbaux.

5. ESG est un analyseur à large couverture écrit en PROLOG

6. Le corpus reader's Digest regroupe l'ensemble des articles du magazine Reader's Digest parus entre 1993 et 1996.

L'auteur s'est intéressé à deux aspects verbaux : la stativité qui est représentée par les verbes d'état et la télélicité<sup>7</sup>. L'auteur s'est basé sur quatorze marqueurs pour déduire les catégories aspectuelles des propositions qui admettent ces marqueurs. Par exemple, une phrase qui apparaît dans une forme progressive 3.4a (admet le marqueur du progressif) représente une action qui est en train de se dérouler (extended event). Une phrase qui admet un adverbial de délimitation temporelle (in-PP) représente une action qui se termine (culminated event) ou possède un aspect télélique 3.4b.

(3.4) (a) action progressive :

*she was prospering in Europe*

(elle prospérait en Europe)

(b) action télélique :

*Derek built his house in one year*

(Derek a construit sa maison en une année)

### 3.3.4 La classification sémantique des verbes

Les variations de diathèse des verbes sont utilisées dans la classification sémantique des verbes. Levin [1993] a été la première à exploiter dans ses travaux le lien qui existe entre les variations de diathèse et les sens des verbes. Elle a stipulé que les verbes ayant les mêmes variations de diathèse - variations dans la réalisation de leurs structures d'arguments - sont supposés partager certains composants sémantiques et peuvent être organisés dans des classes sémantiques cohérentes [Lapata and Brew, 1999]. Le travail de Levin sur les variations de diathèse a influencé de nombreux travaux : désambiguïsation des sens des mots [Dorr and Jones, 1996], traduction automatique [Dang et al., 1998], acquisition automatique des lexiques [McCarthy and Korhonen, 1998, Schulte im Walde, 1998], évaluation [Stevenson and Merlo, 1999], etc. Des travaux de classifications ont même été effectués dans d'autres langues (bengali, allemand, anglais et coréen) [Jones et al., 1994]. D'autres travaux ont développé des méthodes de classification verbale basées sur d'autres propriétés. Par exemple, Merlo and Stevenson [2001] ont décrit une méthode de classification verbale automatique qui repose sur les rôles thématiques qui ont été attribués aux participants.

---

7. Ce terme désigne l'aspect du verbe et fait référence à un événement dont l'activité possède un point de culmination. Par exemple, la phrase *I made a fire* (j'ai fait un feu) comporte un point de culmination qui représente un nouvel état (l'allumage du feu). Les verbes qui possèdent cette propriété sont dits verbes téléliques *fall* (tomber), *kick* (tirer), *make (something)* (faire (quelque chose))

### 3.3.5 La compréhension de textes multilingues

Aone and McKee [1996] ont construit automatiquement des lexiques contenant des représentations conceptuelles des correspondances prédicats-arguments à partir de textes multilingues (Anglais, Espagnol et Japonais). Ces correspondances ont été extraites automatiquement. Pour pouvoir établir ces dernières, les auteurs ont classifié les verbes suivant leurs types de situations où le type de situation représente la propriété aspectuelle du verbe ainsi que les différents rôles thématiques associés aux arguments. Par exemple, le verbe *die* (mourir) est associé au type PROCESS-OR-STATE (Il décrit un processus ou un état) et admet un seul argument avec le rôle thématique (thème). Le type de situation se caractérise par le fait qu'il est indépendant du langage, c'est-à-dire, dans n'importe quelle langue, un verbe admet les mêmes rôles thématiques. Par exemple, dans les trois langues, le verbe *die* admet toujours un seul argument dans le rôle d'un thème. Cette propriété permet d'éviter de définir des règles d'association sémantiques pour chaque verbe lorsqu'il est présent dans une langue donnée. L'étape suivante a consisté à associer à chaque rôle thématique qui appartient à un type de situation donné, la fonction grammaticale (sujet, objet, etc.) qui lui correspond.

## 3.4 Les liens lexicaux entre les différents prédicats

Beaucoup d'applications en TAL nécessitent de connaître les liens lexicaux qui existent entre les items lexicaux en général et les différents types de prédicats en particuliers. Découvrir ces liens permet de déduire les propriétés communes que partagent ces prédicats. Selon Cruse [1986], les liens lexicaux sont soit des relations hiérarchiques basées sur la hiérarchie taxinomique (hyperonymie, hyponymie, et l'implication (entailment)) ou des relations de congruence non hiérarchique (identity, overlap, synonymie, antonymie). Les liens lexicaux sont de différents types : morphologiques, basés sur les variations catégorielles, morpho-sémantiques ou basés sur les fonctions lexicales. Tous ces types de liens ont servi de base pour construire différents lexiques : WordNet [Fellbaum, 1998], HowNet [Dong, 2000], CatVar [Habash and Dorr, 2003], etc.

### 3.4.1 Les liens morphologiques

La morphologie est un domaine de la linguistique qui s'intéresse à la structure interne des mots. On la divise traditionnellement en morphologie flexionnelle et morphologie dérivationnelle. La première décrit les changements que peut subir un mot lorsqu'on lui adjoint des désinences dans le but d'exprimer des catégories grammaticales (le nombre, le genre, la personne, etc.) ou des catégories sémantiques (animé, comptable, etc.). Ces changements n'ont

aucune incidence sur la catégorie syntaxique du mot, par exemple, la pluralisation d'un nom en fera toujours un nom. À la différence de la flexion, la dérivation peut affecter la catégorie syntaxique d'un mot ou son sens en produisant un nouveau mot. Par exemple, *nation* est un nom alors que *national* peut être un adjectif.

Le plus souvent, les ressources lexicales, ne s'intéressent pas aux liens morphologiques qui existent entre les différents items lexicaux car, très souvent, la morphologie est jugée régulière et les affixes ne sont, dans la plupart des cas, la cause d'aucune information ambiguë concernant la catégorie syntaxique ou le sens du mot dérivé [Fellbaum et al., 2007]. Or, connaître les différentes formes d'un mot à travers l'établissement des différents liens morphologiques qui le lient aux autres mots est utile pour l'efficacité des systèmes de recherche. Il y a un lien morphologique entre deux items lexicaux lorsqu'il existe une relation de dérivation entre ces deux items. Les liens morphologiques présentent un grand intérêt car ils sont le point de départ pour déduire les liens syntaxiques et/ou sémantiques qui existent entre les mots et leurs formes dérivées. Par exemple, à partir des deux phrases *the Government manipulated the election* (Le Gouvernement a manipulé les élections) et *the manipulation of the election by the Government* (La manipulation des élections par le Gouvernement), on remarque que le prédicat *manipulate*, dont la structure argumentale est *manipulate(Government, election)*, est lié morphologiquement avec le nom *manipulation*. À partir de ce lien, on peut déduire que : *manipulation* joue le rôle d'un prédicat qui possède la même structure argumentale que *manipulate*, *manipulation(Government, election)*. Habash and Dorr [2003] ont aussi montré l'importance des liens morphologiques en les utilisant dans la construction de CatVar (Categorical Variations), un lexique basé sur les variations catégorielles (voir le paragraphe qui concerne les variations catégorielles). Ce lexique a été utilisé dans de nombreuses applications en TAL : Generation-Heavy Hybrid Machine Translation<sup>8</sup> (GHMT) [Habash, 2002], Headline Generation<sup>9</sup> (HeadGen) [Zajic et al., 2002]. Pour vérifier l'existence des liens morphologiques ou d'en obtenir de nouveaux, différentes méthodes sont utilisées : des méthodes qui utilisent les transducteurs, des méthodes qui utilisent les variations catégorielles, des méthodes qui utilisent des algorithmes d'approximations (Réductionniste, Expansionniste), etc.

**1) Les liens morphologiques dans NOMLEX-PLUS :** Comme nous l'avons vu précédemment (Cf. Chapitre 2.4), NOMLEX-PLUS est un lexique qui décrit les propriétés

---

8. Selon les auteurs, Generation-Heavy Hybrid Machine Translation est une méthode de traduction entre deux langues qui ont des structures divergentes et dotées de ressources asymétriques, c'est-à-dire, une langue source pauvre en ressources linguistiques et une langue cible riche en ressources, tels que l'Espagnol et l'Anglais

9. HeadGen est une application basée sur les chaînes cachées de Markov qui permet de générer automatiquement des titres pour les articles journalistiques. Les auteurs voulaient générer des titres plus informatifs que les titres qui ont été créés par leurs propres auteurs.

syntaxiques des structures argumentales associées à différents prédicats nominaux. Pour enrichir NOMLEX et créer NOMLEX-PLUS, Meyers et al. [2004a] ont utilisé des méthodes semi-automatiques qui ont pour fonction d'extraire les liens morphologiques qui existent entre différentes unités lexicales appartenant à des catégories syntaxiques distinctes (nominalisation/verbe, adjectif/adverbe, verbe/nom et nominalisation/adjectif). L'enrichissement de NOMLEX s'est déroulé en plusieurs étapes :

1. Les auteurs se sont d'abord intéressés à l'extraction des couples nominalisation/verbe. Pour cela, ils ont utilisé une méthode qui extrait à partir de COMLEX (Cf. Chapitre 2.4) tous les couples qui partagent les racines les plus longues. Par exemple, les mots du couple *abrade/abrasion* (abraser/abrasion) ont en commun la racine *abra-*.
2. La même méthode a été employée pour extraire les couples adjectif/adverbe qui ont des liens morphologiques entre eux. Ces couples ont été insérés dans ADJADV - un dictionnaire qui définit les emplois adverbiaux des adjectifs et qui est utilisé dans leurs travaux sur NomBank. Par exemple le mot *fine* (bien, beau, excellent, etc.) possède le même sens, qu'il soit adjectif (*fine behavior*, beau comportement) ou adverbe (*he behaved fine*, il s'est bien comporté).
3. Tous les couples verbe/nom qui ont des racines communes et qui possèdent des paires de suffixes spécifiques sont extraits. Par exemple, le couple *anesthetize/anesthetist* (anesthésier/anesthésiste) partage la racine *anesthet-* et le couple de suffixe *-ize/-ist*.
4. L'extraction des couples de nominalisation/adjectif repose aussi sur les liens morphologiques mais la méthode adoptée diffère quelque peu des deux premières. Pour cela, à partir du Penn Tree Bank, ont été classés manuellement des couples de nominalisations associées aux adjectifs qui leur correspondent. En analysant les suffixes de chaque couple, ils ont mis en évidence une certaine régularité qui est exploitée pour dériver de nouveaux couples. Par exemple, le couple de suffixes *-ability/-able*, extrait à partir des items lexicaux *durability/durable* (durabilité/durable), a servi à identifier de nouveaux couples *availability/available* (disponibilité/disponible), *absorbability/absorbable*, etc..

Ces méthodes morphologiques ont permis de construire un lexique qui décrit les propriétés syntaxiques des structures argumentales de 5 450 nouvelles nominalisations (4 900 verbales et 550 adjectivales), de 1 600 prédicats nominaux (noms partitifs, relationnels, etc.), auquel se rajoutent les 1 000 nominalisations verbales qui appartiennent déjà à NOMLEX.

**2) Les variations catégorielles et les algorithmes d'approximation :** La variation catégorielle d'un mot qui appartient à une catégorie grammaticale donnée est un autre mot qui possède un lien dérivationnel avec ce premier mot et dont la catégorie grammaticale peut être

différente. Par exemple, les variations catégorielles du verbe *abhor* (abhorrer) sont : le nom *abhorrence* (aversion) et l'adjectif *abhorrent* (odieux). L'utilisation des variations catégorielles permet de relier les mots qui possèdent des liens morphologiques et dérivationnels tout en spécifiant les catégories syntaxiques de chaque mot. En raison des nombreuses variations catégorielles qui existent dans différentes langues, la construction de ressources basées sur les variations catégorielles est très utile dans de nombreuses applications en TAL (traduction automatique, recherche d'informations et construction de lexiques). Ainsi :

- en traduction automatique : Habash et al. [2002] déclarent que 98% des divergences de traduction (ces divergences sont définies comme des variations dans la structure sémantique des langages source et cible) impliquent certaines formes de variations catégorielles ;
- en recherche d'informations : certains systèmes ont besoin de réduire les différentes variantes d'un mot à une racine commune pour améliorer la reconnaissance des modèles (pattern matching) [Xu and Croft, 1998].

Pour construire automatiquement des ressources qui se basent sur les variations catégorielles, certains travaux ont utilisé des méthodes qui s'appuient sur des algorithmes d'approximation. Ces algorithmes utilisent deux types d'approximation : (i) approximation Réductionniste (Analytique) et (ii) approximation Expansionniste (Générative). Le premier de ces algorithmes a pour rôle de transformer les formes de surface des différents items lexicaux en leurs formes radicales - le racinisateur (stemmer) de Porter [1997] utilise ce type d'approximation. Le second de ces algorithmes, reposant sur l'approche Expansionniste, utilise des règles de dérivation pour surgénérer de nouvelles entrées et applique ensuite des modèles statistiques pour effectuer un classement ou une sélection parmi les mots générés - le générateur morphologique Nitrogen [Langkilde and Knight, 1998] exploite ce type d'approximation. Néanmoins, selon Habash and Dorr [2003], les algorithmes de ces deux méthodes présentent les problèmes suivants :

- La sous-racinisation/sur-racinisation et la sous-génération/sur-génération génèrent des données qui manquent de qualité à cause de leur nature d'approximation brute, ce qui a pour conséquence un manque d'efficacité. Par exemple, le racinisateur de Porter va relier les mots *commune<sub>N</sub>*, *communication<sub>N</sub>* et *communism<sub>N</sub>* à la racine *commun*, mais il n'établit pas le même lien pour les mots *communist<sub>N</sub>* et *communicable<sub>A</sub>* qui sont respectivement reliés aux racines *communist* et *communic*.
- Le générateur morphologique Nitrogen va associer le mot *develop* (développer) à une dizaine de mots y compris *developage*, *developication* et *developy*, ce qui est faux. Il n'y a que deux associations qui sont correctes parmi tous ces liens (*development* et *develop-*

*ing*). De telles sur-génération vont induire une expansion exponentielle de l'espace de recherche et ramener du bruit dans l'algorithme de recherche, ou même, laisser apparaître des mots erronés dans le haut des listes proposées.

- Les deux approximations sont unidirectionnelles, c'est-à-dire que le racinisateur ne peut pas être utilisé pour la génération et le sur-générateur morphologique ne peut pas être utilisé comme un racinisateur.
- La racinisation ne peut pas traiter la différence produite par les différents sens d'un mot donné. Par exemple, le mot *gravitation<sub>N</sub>* (gravitation) va être relié à la racine *gravity<sub>N</sub>* qui possède le sens de *serious* (sérieux) au lieu d'être relié au sens *force-of-gravity* (force de gravité) de la même racine [Krovetz, 1993].

**3) CatVar :** CatVar [Habash and Dorr, 2003] est une ressource lexicale basée sur les variations catégorielles. En tant que ressource, elle a été créée à partir des deux approches avec la particularité de diminuer le bruit dû à la sous/sur-racinisation/génération. La version 2.0 de CatVar inclut 62 232 groupes couvrant 96 368 lexèmes uniques. 62% des lexèmes sont des noms, 24% des adjectifs, 10% des verbes et 4% sont des adverbes. Pour déterminer la précision et le rappel de CatVar, les auteurs ont demandé à 8 locuteurs natifs d'évaluer 400 groupes dont la sélection était aléatoire. L'évaluation a permis de mesurer une précision de 91.82% et un rappel de 81.00%.

Cette ressource a été utilisée dans de nombreuses applications monolingues et multilingues (traduction automatique, recherche d'informations, etc.). Le développement de CatVar a nécessité l'utilisation de plusieurs ressources et algorithmes :

1. Le lexique Englex [Antworth, 1990] qui est un lexique morphologique qui contient 20 000 entrées de différents types : affixes, racines et radicaux indivisibles. Ce lexique est utilisé par les applications de prétraitement morphologique de textes afin de produire un texte morphologiquement annoté. Il peut être utilisé aussi pour explorer la structure morphologique de l'anglais.

Englex a été développé pour être associé avec KIMMO [Antworth, 1990]. Il a pour fonction de générer et/ou d'analyser des mots en utilisant un modèle à deux niveaux [Koskeniemi, 1984] (two-level model). Il s'agit d'un modèle réalisé à l'aide d'automates à états finis (transducteurs). Il propose un cadre indépendant du langage pour décrire les phénomènes morphologiques et phonologiques associés à la flexion, à la dérivation et à la composition des mots. Selon Koskeniemi and Church [1988], le modèle possède deux aspects intéressants : (i) c'est un formalisme linguistique pour décrire les phénomènes

phonologiques et (ii) c'est un outil informatique capable d'implémenter les descriptions d'un langage donné dans le but d'avoir un système opérationnel capable de reconnaître les différentes formes des mots et de les générer. Ce modèle se compose de trois représentations (morphologique, lexicale ou morphophonologique et de surface), reliées par 2 systèmes (le lexique et les règles phonologiques). Il permet de faire la distinction entre le niveau lexical et le niveau de surface. Le premier niveau traite les mots lexicaux (représentations lexicales), qui proviennent de la concaténation des morphèmes (représentation morphologique) contenus dans le lexique ou le dictionnaire. Par exemple, *fox+s* est une forme lexicale qui provient de la concaténation des deux morphèmes *fox* et *+s* (où le + indique la présence d'un affixe). Le second niveau traite les mots de surface (représentation de surface), tels qu'ils apparaissent dans les textes. Par exemple, le mot de surface associé au mot lexical *fox+s* est *foxes* (renards). Les règles à deux niveaux ont pour rôle de comparer les divergences qui existent entre les représentations associées à ces deux niveaux et de les traiter.

2. Les lexiques NOMLEX et LDOCE pour extraire tous les liens de base qui existent entre les différents mots.
3. Le lexique de Brown a été utilisé pour avoir une large couverture de lexèmes. Pour cela, a été employé l'analyseur morphologique d'Englex utilisant les catégories syntaxiques qui figurent dans le Penn Tree Bank.
4. Le racinisateur de Porter a été finalement utilisé dans l'étape de regroupement (clustering) pour augmenter les liens de base et créer des groupes contenant les mots qui sont des variations catégorielles entre eux *hunger<sub>N</sub>*, *hungry<sub>A</sub>*, *hungry<sub>V</sub>*, *hungriness<sub>N</sub>*.

### 3.4.2 Les liens morpho-sémantiques

WordNet est un lexique qui a été initialement conçu pour imiter le modèle sémantique de la mémoire humaine où les connaissances lexicales d'un locuteur sont représentées par un réseau dans lequel les mots et les concepts possédant les mêmes sens sont reliés entre eux. Les concepteurs de WordNet ont ensuite voulu accroître le nombre de connexions de ce réseau en ajoutant les mots qui sont morphologiquement et sémantiquement reliés [Fellbaum and Miller, 2003].

L'anglais est une langue qui possède des règles flexionnelles assez simples. Une liste de règles associée à une autre liste d'exceptions suffisent à obtenir le pluriel de la plupart des mots. WordNet incorpore un programme qui permet d'obtenir les flexions des mots réguliers et irréguliers de façon satisfaisante. Par contre, la morphologie dérivationnelle est beaucoup plus compliquée à incorporer dans un système automatique. Le recours à ce type de morphologie

est particulièrement utilisé dans le domaine de la traduction automatique. Ceci explique l'importance qu'il y a à bien identifier et à relier les mots qui sont sémantiquement liés, même s'ils appartiennent à des catégories syntaxiques différentes. Par exemple, le verbe *digest* (digérer) et sa nominalisation *digestion* possèdent deux sens :

1. un sens physiologique qui signifie "transformer la nourriture en des aliments qui peuvent être absorbés par l'organisme". Ce sens est illustré par l'exemple *she digested her dinner* (elle a digéré son dîner). Dans ce cas, la nominalisation *digestion* possède aussi le même sens physiologique : *he consulted the doctor about his digestion* (il a consulté le médecin à propos de sa digestion) ;
2. un sens psychologique ou mental qui signifie "absorber ou assimiler mentalement". L'exemple *they digested the information* (ils ont assimilé l'information) montre cet autre sens. Là encore, la nominalisation *digestion* possède le sens : *his appetite for facts was better than his digestion* (Son intérêt pour les faits était meilleur que son assimilation).

N'importe quel utilisateur d'un dictionnaire comprendra implicitement que le sens physiologique de *digestion* est dérivé directement du sens physiologique du verbe *digest* (la même remarque s'applique pour le sens psychologique/mental), mais un système automatique ne fera pas cette déduction, à moins de lui dire explicitement quel sens il doit prendre. Ceci montre clairement que les liens morphologiques ne suffisent pas à eux seuls à régler ces problèmes d'ambiguïté. Ceci a motivé Fellbaum and Miller [2003] pour qu'ils introduisent dans WordNet des liens morpho-sémantiques. Ce type de liens est la combinaison de deux types de liens : les liens sémantiques qui permettent de connecter les sens et les liens morphologiques qui sont déjà présents dans les versions antérieures de ce lexique.

Les liens morpho-sémantiques sont basés sur le principe qu'on peut relier sémantiquement certains mots en se basant sur le type du suffixe ajouté. Certains suffixes sont associés à des sémantiques bien spécifiques et lorsqu'ils sont rattachés à des racines, ils produisent des mots dérivés dont le sens est sémantiquement relié à celui de leurs formes de base. Par exemple, Aronoff [1976] (page 48) déclare qu'un des sens du suffixe *-able* lorsqu'il est adjoint à un verbe est "facile à X" où X est le verbe de base. Cette règle est productive car elle permet d'extraire les mots qui sont sémantiquement reliés. Par exemple, les adjectifs *readable* et *learnable* sont morphologiquement et sémantiquement reliés aux verbes *to read* (lire) et *to learn* (apprendre) car leurs sens respectifs sont "facile à lire" et "facile à apprendre". De même, le suffixe *-ness* indique "un état ou une condition dénoté par l'adjectif de base". Ceci permet d'établir des liens sémantiques entre les adjectifs et les adverbes qui sont morphologiquement reliés. Par exemple,

*happiness* (bonheur) reflète un état dénoté par l'adjectif *happy* (heureux). Il est facile pour un humain d'utiliser ces règles dans le but de trouver les liens morpho-sémantiques qui relient les différents mots, par contre, elles ne sont pas faciles à exploiter par les systèmes automatiques pour les raisons suivantes :

- certains suffixes sont polysémiques. Par exemple, le suffixe *-able*, représenté par l'adjectif *fashionable* possède un deuxième sens qui est : "caractérisé par X" où X est la forme de base [Aronoff, 1976](page 48) ;
- ces différentes règles ne sont pas absolues. Selon Aronoff [1976], le suffixe *-ee* est associé uniquement aux verbes transitifs (par exemple, *payee* (celui qui est payé) et *employee* (celui qui est employé)), mais cette règle n'est pas correcte, comme le montre l'exemple *standee* (celui qui attend) qui est associé au verbe *to stand* (attendre) qui peut être intransitif.

Pour pouvoir inclure ce type de liens dans WordNet, les concepteurs de ce lexique ont suivi plusieurs étapes :

1. la première étape a consisté à identifier tous les mots qui sont reliés entre eux et à construire une liste de noms et de verbes qui possèdent des liens dérivationnels en se basant sur un ensemble de suffixes dérivationnels ;
2. ils ont ensuite identifié, à partir de WordNet, tous les noms qui se terminent par les suffixes : *-acy, -age, -al, -ance, -ancy, -ant, -ard, -ary, -ate, -ation, -ee, -er, -ery, -ing, -ion, -ure* et les verbes qui se terminent par : *-ate, -ify, -ize* ;
3. la troisième étape a consisté à construire une liste qui regroupe tous les couples nom/verbe qui sont des homographes (qui n'ont pas d'affixes), par exemple, le verbe *to bow* (courber, incliner) et le nom *bow* (avant, proue), ainsi que des noms-verbes qui sont proches des homographes mais qui impliquent des particules (*run off/runoff* (s'écouler/trop-plein), *put down/put-down* (poser/refus brutal)).

Ils ont ainsi créé vingt-quatre listes où chacune d'elles regroupe les paires des noms-verbes qui sont reliés par un affixe donné. Pour faire face aux problèmes cités ci-dessous concernant la difficulté de relier automatiquement les sens des verbes et des noms, ils ont créé une interface graphique qui affiche pour chaque paire, tous les sens pris par les verbes et leurs nominalisations, puis ils ont fait appel à des humains pour qu'ils établissent ces relations. Cette démarche a permis d'accroître la connectivité de WordNet en créant 21 000 nouveaux liens morpho-sémantiques, ce qui est bénéfique pour beaucoup d'applications en TAL

### 3.4.3 Conclusion

Nous avons vu dans ce chapitre que les structures prédicatives facilitent la capture des régularisations syntaxiques en permettant de trouver une représentation commune (canonique) à plusieurs formes syntaxiques (non canoniques). Ces dernières sont dues au fait que le langage naturel dispose de plusieurs réalisations syntaxiques, telles que l'actif, le passif, le gérondif, l'alternance du datif, etc. Elles sont utilisées principalement dans (i) l'acquisition automatique des cadres de sous-catégorisations, des contraintes de sélection et des informations sémantiques, (ii) dans la classification sémantique des verbes et (iii) dans la compréhension de textes multilingues.

Les structures prédicatives intègrent différents types de prédicats : verbaux, nominaux, adjectivaux, etc. Ces derniers partagent un ensemble de propriétés lexicales, syntaxiques et sémantiques. La mise en évidence de ces propriétés a reposé sur différents types d'approches qui ont permis de construire les ressources électroniques qu'on a vues : (i) l'approche syntaxique avec substrat morphologique qui se base sur l'étude des liens morphologiques entre les prédicats. Cette approche a conduit à la construction de différentes ressources, telles que : COMLEX, NOMLEX, NOMLEX-PLUS, etc., (ii) l'approche fortement morphologique qui se base sur les variations catégorielles et qui a servi dans la construction de CatVar et (iii) l'approche morpho-sémantique, qui est basée sur l'étude des propriétés morphologiques et sémantiques qui existent entre les différents prédicats. Cette approche a servi dans la construction de WordNet, FrameNet et VerbNet.

On a également vu que la plupart des travaux qui ont concernés les structures prédicatives se sont intéressés aux prédicats verbaux. Dans le chapitre suivant, l'étude des structures argumentales prédicatives associées aux schémas nominaux nous permettra de trouver une représentation commune de la structure syntaxique d'un groupe nominal prédicatif (GNpréd) à partir de plusieurs représentations nominales équivalentes. Ils nous permettront également d'associer un seul GNpréd à différents schémas, ce qui facilitera la capture et l'acquisition des arguments nominaux.

## **Chapitre 4**

# **Acquisition des Structures Argumentales pour les Prédicats Nominaux**

## 4.1 Introduction

L'arité des arguments d'une nominalisation ainsi que les fonctions grammaticales des arguments qu'elle admet dépendent directement de la classe grammaticale du prédicat verbal (transitif, intransitif, etc.) dont elle dérive. Connaître les liens qui existent entre les différents types de schémas verbaux et la façon dont leurs groupes nominaux prédicatifs (GNpréd) se construisent facilite l'acquisition des différents arguments nominaux. Cette information est obtenue par l'analyse des schémas de complémentation des verbes et des nominalisations. Pour pouvoir analyser une grande quantité de schémas et les manipuler aisément, nous avons développé PredicateDB [Malik and Royauté, 2009, 2007], une plate-forme se composant d'un ensemble d'outils et d'une base de données qui nous permet de rapprocher les propriétés des deux types de schémas. Cette plate-forme nous a permis également d'extraire les nominalisations suivant leur comportement syntaxique dans le but de construire semi-automatiquement un lexique de nominalisations pour un analyseur syntaxique et sa grammaire de GNpréd [Royauté et al., 2006, 2007].

## 4.2 Le groupe nominal

Un groupe nominal est un ensemble de constituants reliés à un nom de tête qui représente le noyau du groupe et qui peut être réduit à ce seul nom. Un groupe nominal (GN) peut être prédicatif ou non prédicatif suivant le type du nom de tête.

### 4.2.1 Structure distributionnelle du groupe nominal

La structure distributionnelle d'un GN se compose principalement, comme le montre la figure 4.1, d'un nom de tête autour duquel se regroupent différents éléments : un ou plusieurs déterminants et un ou plusieurs modificateurs. Ces derniers sont dits prémodificateurs quand ils précèdent le nom de tête et postmodificateurs lorsqu'ils le suivent. L'exemple 4.1<sup>1</sup> montre un groupe nominal composé d'un déterminant, d'un prémodificateur, d'un nom de tête et d'un postmodificateur. Il est intéressant de noter que tous les éléments (sauf le nom de tête) d'un groupe nominal ont la particularité d'être optionnels et peuvent donc être effacés. Ainsi, le GN de notre exemple peut être réduit au seul nom *days* (les éléments optionnels sont mis entre crochets).

(4.1) *[all those]<sub>Dét</sub> [fine warm]<sub>PréMod</sub> [days]<sub>tête</sub> [in the country last year]<sub>PostMod</sub>*

1. La plupart des phrases et des GNpréd qui nous ont servis d'exemples dans ce chapitre ont été extraits principalement à partir des trois ressources suivantes : le site officiel du Gouvernement du Canada ([canada.gc.ca](http://canada.gc.ca)) qui met à la disposition de l'utilisateur des traductions anglais-français, le web et le livre de Quirk et al. [1987]

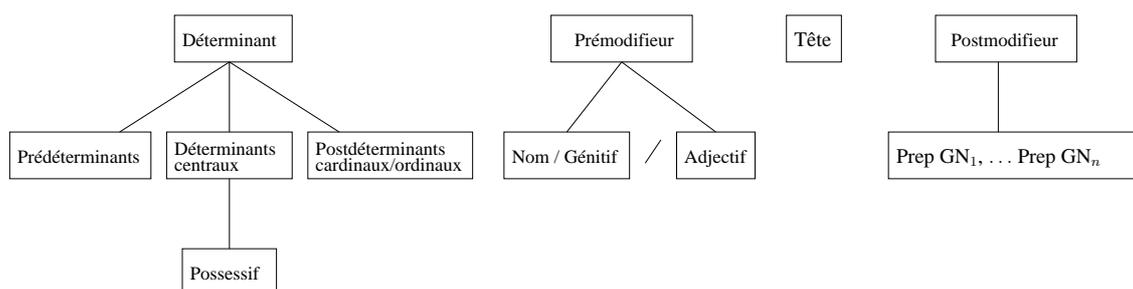


FIGURE 4.1 – Le schéma d'un groupe nominal

(toutes ces belles journées chaudes dans le pays l'année dernière)

1. **Le déterminant** : Les déterminants sont des constituants du GN qui se placent avant le nom de tête dans le but de l'actualiser, c'est-à-dire lui attacher différentes propriétés : singulier, pluriel, cardinalité, défini, indéfini, possession, etc. Nous distinguons trois classes de déterminants : prédéterminants, déterminants centraux et postdéterminants. Le groupe nominal *all these last few days* (tous ces derniers jours) admet un déterminant central (*these*) qui est précédé par un prédéterminant (*all*) et suivi par un postdéterminant (*few*) :

- les déterminants centraux : les plus communs sont les articles *the* et *a/an* (exemples 4.2a et 4.2b), mais il existe d'autres déterminants tels que les démonstratifs (*this, those, etc.*, exemple 4.2c), les quantifieurs (*some, any, etc.*, exemple 4.2d), etc. Les déterminants centraux sont mutuellement exclusifs, c'est-à-dire que deux déterminants centraux ne peuvent pas figurer dans le même groupe nominal (exemple 4.2e).

- (4.2) (a)  $[the]_{DétCent} book$   
 (le livre)  
 (b)  $[a]_{DétCent} boat$   
 (un bateau)  
 (c)  $[this]_{DétCent} new bicycle$   
 (cette nouvelle bicyclette)  
 (d)  $[some]_{DétCent} water$   
 (\*de l'eau)

- (e) *\*a the boat*  
(un le bateau)

Les adjectifs possessifs (*my, your, their, etc.*), nommés aussi déterminants possessifs, sont un autre type de déterminants. Ils sont considérés comme des déterminants centraux car, dans la plupart des cas, ils possèdent le même comportement et les mêmes contraintes que ces derniers : ils précèdent le nom de tête (exemple 4.3a), ils peuvent être précédés par des prédéterminants (exemple 4.3b), suivis par des postdéterminants (exemple 4.3c). Les adjectifs possessifs sont mutuellement exclusifs (exemple 4.3d), excepté lorsqu'ils sont associés à *own* dans une construction emphatique et déterminative (exemple 4.3e).

- (4.3) (a) *[her]<sub>DétPoss</sub> passion for the theater*  
(sa passion pour le théâtre)  
(b) *[all]<sub>PréDet</sub> [my]<sub>DétPoss</sub> books*  
(tous mes livres)  
(c) *[his]<sub>DétPoss</sub> [two]<sub>PostDét</sub> children*  
(ses deux enfants)  
(d) *\*my their car*  
(\*ma leur voiture)  
(e) *Michelle cooks [her]<sub>DétPoss</sub> own dinner every evening*  
(Michelle prépare son propre diner chaque soir)

- les prédéterminants : Ces déterminants précèdent toujours les déterminants centraux. Ils peuvent prendre différentes valeurs : *all, both, half, etc.* (exemples 4.4a à 4.4c). Comme les déterminants centraux, les prédéterminants sont mutuellement exclusifs (exemple 4.4d).

- (4.4) (a) *[all]<sub>PréDet</sub> arrangements of a set*  
(toutes les dispositions d'un ensemble)  
(b) *[both]<sub>PréDet</sub> musicians in the band*  
(les deux musiciens de la bande)  
(c) *[half]<sub>PréDet</sub> [an]<sub>DétCent</sub> hour*  
(une demi-heure)  
(d) *\*all both boys*

(\*tous deux garçons)

- les postdéterminants : Ces déterminants se placent après les prédéterminants et les déterminants centraux mais précèdent les prémodificateurs tels que les noms et les adjectifs (exemple 4.5a). Les postdéterminants peuvent être des adjectifs cardinaux : *one, two, three, etc.* (exemple 4.5b), ordinaux : *first, second, etc.* (exemple 4.5c) ainsi que les quantifieurs *few, many, a large number of, etc.* (exemple 4.5d). Certains postdéterminants possèdent des contraintes particulières : (i) dans certains cas, le postdéterminant *one* ne peut pas être précédé par le déterminant central indéfini *a* (exemple 4.5e) mais il est possible de l'employer avec *the* (exemple 4.5f) (ii) À la différence des prédéterminants et déterminants centraux qui sont mutuellement exclusifs, il est possible que deux postmodificateurs figurent dans le même groupe nominal. Dans la plupart des cas, les ordinaux précèdent les cardinaux (exemple 4.5g).

(4.5) (a) *[my]PostDét [new]PréMod office*

(mon nouveau bureau)

(b) *her [three]PostDét children*

(ses trois enfants)

(c) *the [first]PostDét smiley*

(le premier sourire)

(d) *the [many]PostDét uses of the forest*

(les nombreuses utilisations de la forêt)

(e) *\*a one book*

(\*un un livre)

(f) *the one book I like*

(le seul livre que j'aime)

(g) *another three weeks*

(trois semaines encore/de plus)

2. **Le prémodifieur** : Un prémodifieur est un mot qui se place avant le nom de tête et a pour fonction de rajouter une information descriptive à ce dernier. Un prémodifieur peut être de différents types :

- Nom/adjectif : dans le cas où le prémodifieur est un nom, il a pour fonction de préciser le type du nom de tête qu'il précède (exemple 4.6a). S'il est adjectif, il a pour rôle de

dénoter une qualité ou une propriété appartenant au nom de tête (exemple 4.6b).

- (4.6) (a) *the [office]<sub>PreMod</sub> furniture*  
(le mobilier de bureau)  
(b) *all those [beautiful]<sub>PreMod</sub> cars*  
(toutes ces belles voitures)

- Génitif : le génitif peut être prémodifieur. Dans ce cas, il peut exprimer la notion de possession (exemple 4.7a) ou de classification (exemple 4.7b). Dans le premier exemple, *daughter* possède une *robe*, ce qui exprime une notion de possession. Le deuxième exemple montre que le génitif *girl's* représente une certaine catégorie de lycées (pour filles).

- (4.7) (a) *my [daughter's]<sub>Gén</sub> dress*  
(la robe de ma fille)  
(b) *there are several [girl's]<sub>Gén</sub> secondary schools*  
(Il existe de nombreux lycées pour filles)

3. **Le postmodifieur** : Un postmodifieur est un modifieur qui se positionne après la tête qu'il modifie. Un nom de tête peut être modifié par différents types de postmodifieurs : un adjectif (exemple 4.8a), un adverbe (exemple 4.8b), un groupe prépositionnel (exemple 4.8c), une relative (exemple 4.8d) ou une proposition participiale (exemple 4.8e).

- (4.8) (a) *something [different]<sub>PostMod</sub>*  
(quelque chose de différent)  
(b) *the direction [back]<sub>PostMod</sub>*  
(la direction de retour)  
– ((c)] *ticket [to Paris]<sub>PostMod</sub>*  
(billet pour Paris)  
(d) *a painting [that costs a fortune]<sub>PostMod</sub>*  
(un tableau qui coûte une fortune)  
(e) *the man [observed near the scene of the crime]<sub>PostMod</sub>*  
(l'homme observé près de la scène du crime)

## 4.2.2 Le groupe nominal à tête prédicative

Un groupe nominal prédicatif (GNpréd) est une structure prédicative dont le nom de tête possède des arguments. Ce dernier peut être dérivé d'un verbe (exemple 4.9a) où *examination* est la nominalisation associée au verbe *to examine*. Notre travail ne va s'intéresser qu'à ce type de nominalisations. Le nom de tête peut être également dérivé d'un adjectif (exemple 4.10a) [Grefenstette and Teufel, 1995] où *impatience* est le prédicat nominal de l'adjectif *impatient* ou associé à un verbe support comme *make*, *have*, *take*, etc. (exemple 4.11a).

(4.9) (a) *Bryan examined<sub>v</sub> the car*

(Bryan a examiné la voiture)

(b) *the examination<sub>NPred</sub> of the car by Bryan*

(l'examen de la voiture par Bryan)

(4.10) (a) *John was impatient<sub>adj</sub> with Juliette*

(John était impatient avec Juliette)

(b) *the impatience<sub>Npred</sub> of John with Juliette*

(l'impatience de John avec Juliette)

(4.11) (a) *Mrs. Marcos appealed to President Corazon Aquino*

(b) *Mrs. Marcos made<sub>Vsup</sub> an appeal<sub>N</sub> to President Corazon Aquino*

(Mme Marcos a fait appel au Président Corazon Aquino)

Tout schéma verbal dont le verbe admet une nominalisation est associé à un schéma nominal (GNpréd) qui, dans la grande majorité des cas, conserve les mêmes arguments verbaux. On retrouve donc dans la structure du GNpréd, en grande partie, le schéma de complémentation du verbe dont le prédicat nominal est dérivé [Pasero et al., 2004]. Le GNpréd se caractérise par le fait que tous ses arguments sont optionnels. On dit qu'un GNpréd est saturé lorsque tous les arguments verbaux sont présents et non saturé lorsque un ou plusieurs arguments ont été effacés. Les arguments optionnels sont mis entre crochets. Le GNpréd 4.9b, qui est associé à la phrase 4.9a, est saturé car tous les arguments verbaux (*Bryan* et *car*) sont présents. La figure 4.2 représente le schéma général d'un GNpréd. Ce dernier est une extension de la figure 4.1.

Le schéma d'un GNpréd diffère du schéma d'un GN non prédicatif non d'un point de vue distributionnel mais d'un point de vue fonctionnel. Cela veut dire que les formes et les

positions des constituants des deux schémas sont les mêmes. La différence entre les deux schémas réside dans le fait que les constituants d'un GN<sub>préd</sub> représentent dans la plupart des cas des arguments auxquels on peut associer des fonctions syntaxiques, tels que sujet, complément d'objet direct, etc.

1. **les arguments en position déterminant** : les arguments qui occupent cette position prennent la forme d'un déterminant qui réfère ici un argument sujet comme le montre l'exemple 4.12b.

- (4.12) (a) *he assesses the situation*  
 (il contrôle la situation)  
 (b) [*his*]<sub>sujet=poss</sub> *assessment of the situation*  
 (Son contrôle de la situation)

2. **les arguments en position prémodifieur** : Ces arguments précèdent le prédicat nominal et peuvent prendre la forme d'un nom (exemple 4.13b) ou d'un génitif (exemple 4.13d).

- (4.13) (a) *an abscess formed at the bite site*  
 (un abcès s'est formé à l'endroit de la piqûre)  
 (b) [*abscess*]<sub>sujet=PreMod</sub> *formation at the bite site*  
 (La formation d'un abcès à l'endroit de la piqûre)  
 (c) *cells absorb lipid*  
 (les cellules absorbent le lipid)  
 (d) [*Cells'*]<sub>sujet=Gen</sub> *absorption of lipid*  
 (l'absorption des lipides par les cellules)

3. **les arguments en position postmodifieur** : Ces arguments prennent la forme d'un complément prépositionnel introduit par une préposition (exemple 4.14b).

- (4.14) *The absorption [of Cholesterol]*<sub>Arg=(of)N</sub>  
 (L'absorption du Cholestérol)

Ainsi, les arguments nominaux occupent de types de positions : position gauche -avant le prédicat nominal (déterminant et prémodifieur)- et position droite -après le prédicat nominal

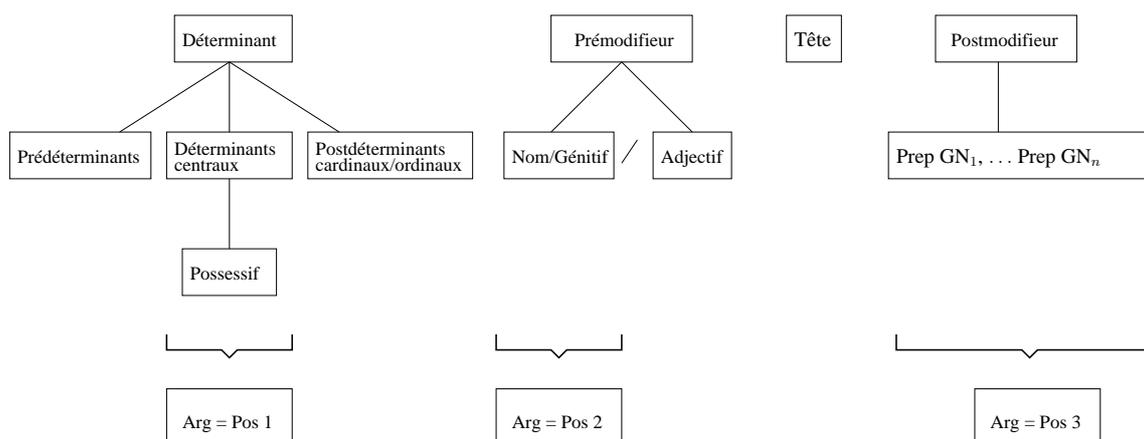


FIGURE 4.2 – Le schéma général d'un groupe nominal prédicatif

(postmodifieur). Ces positions ainsi que le type des fonctions syntaxiques des arguments nominaux représentent le schéma de complémentation du prédicat nominal. Ce dernier dépend du schéma de complémentation du verbe à partir duquel a été dérivée la nominalisation. Chaque type de schéma verbal est associé à un GNpréd particulier dans lequel les arguments nominaux possèdent certaines contraintes sur leurs positions dans le GNpréd, leurs fonctions syntaxiques, etc.

### 4.2.3 Les formes équivalentes du groupe nominal prédicatif

Les arguments nominaux peuvent apparaître sous différentes constructions, ce qui permet à un même groupe nominal prédicatif d'avoir différentes formes distributionnelles équivalentes. Un argument nominal peut être en position prémodifieur sous la forme d'un génitif ou d'un nom ou en position postmodifieur introduit par une préposition spécifique. La capacité que possède un argument d'occuper plusieurs positions est due au fait qu'il existe des équivalences de sens entre les différentes constructions.

#### Équivalences entre prémodifieurs et constructions prépositionnelles

Un GNpréd qui admet un argument en position prémodifieur, sous la forme d'un génitif ou d'un nom, possède dans la plupart des cas une forme équivalente dans laquelle le même argument est en position postmodifieur introduit par une préposition, principalement *of* et *by*.

Un schéma verbal qui n'admet pas de COD (intransitif, transitif indirect, etc.) est associé à un GNpréd dans lequel le sujet est introduit par la préposition *of* (Cf. Section 4.2.2). Ce même sujet peut dans la plupart des cas changer de position et se mettre en position prémodifieur en prenant la forme d'un génitif ou d'un nom. Par exemple, la phrase 4.15a correspond à un schéma intransitif dans lequel *body* joue le rôle du sujet. Cette phrase est associée à plusieurs GNpréd équivalents dans lesquels le sujet *body* se trouve dans différentes positions : en position postmodifieur introduit par la préposition *of* (exemple 4.15b) ou en position prémodifieur sous la forme d'un génitif (exemple 4.15c) ou nom modifieur de nom (exemple 4.15d). Ainsi, dans la plupart des cas, on peut passer d'une construction postmodifieur à une construction prémodifieur sans altérer le sens. Cette équivalence s'applique aussi sur les schémas nominaux avec COD. Les deux arguments de ce schéma que sont le COD et le sujet peuvent être en position postmodifieur introduits respectivement par les prépositions *of* et *by* ou en position prémodifieur en prenant la forme d'un génitif ou d'un nom. Néanmoins, un COD en position postmodifieur introduit par la préposition *of* a plus de chance de se retrouver en position prémodifieur sous la forme d'un nom que d'un génitif [Reeves et al., 1999]. Par exemple, le GNpréd 4.16b qui est associé à une phrase transitive directe (exemple 4.16a) admet un sujet (*tissue*) introduit par la préposition *by* et un COD (*liquid*) introduit par *of*. Ce GNpréd peut avoir plusieurs formes équivalentes dans lesquelles le sujet *tissue* peut être en position prémodifieur en prenant la forme d'un génitif ou d'un nom (exemple 4.16c et 4.16d). Le COD peut lui aussi passer d'une position postmodifieur introduit par la préposition *of* à une position prémodifieur en prenant la forme d'un nom modifieur (exemple 4.16e) et, moins fréquemment, la forme d'un génitif (exemple 4.16f).

(4.15) (a) *the body reacted*

(le corps a réagi)

(b) *the [reaction]<sub>Préd</sub> of the [body]<sub>sujet</sub>*

(c) *the [body's]<sub>sujet</sub> [reaction]<sub>Préd</sub>*

(d) *the [body]<sub>sujet</sub> [reaction]<sub>Préd</sub>*

(la réaction du corps)

(4.16) (a) *the tissue absorbs the liquid*

(le tissu absorbe le liquide)

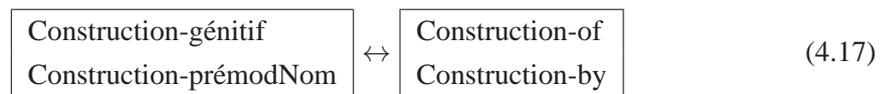
(b) *the [absorption]<sub>Préd</sub> of the [liquid]<sub>Cod</sub> by the [tissue]<sub>Sujet</sub>*

(c) *[tissue's]<sub>sujet</sub> absorption of the liquid*

(d) *[tissue]<sub>sujet</sub> absorption of the liquid*

- (e) *the [liquid]<sub>Cod</sub> absorption by the tissue*  
 (f) *the [liquid's]<sub>Cod</sub> absorption*  
 (l'absorption du liquide par le tissu)

L'équivalence qui existe entre la position postmodifieur et prémodifieur est représentée par le schéma 4.17.

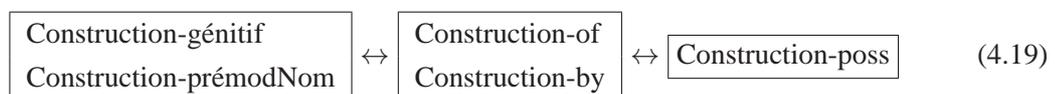


### Extension aux possessifs

Un argument nominal en position postmodifieur introduit par les prépositions *of* et *by* peut dans la plupart des cas prendre la forme d'un possessif en position déterminant. Par exemple, à partir du GNpréd 4.16b, on peut obtenir des GNpréd équivalents dans lesquels le sujet et le COD sont remplacés par des possessifs. Les GNpréd 4.18a et 4.18b montrent que le possessif *its* fait référence respectivement au sujet *tissue* et au COD *liquid*.

- (4.18) (a) *[its]<sub>Sujet</sub> absorption of the [liquid]<sub>Cod</sub>*  
 (Son absorption du liquide)  
 (b) *[its]<sub>Cod</sub> absorption by the [tissue]<sub>Sujet</sub>*  
 (son absorption par le tissu)

À partir du schéma 4.17 et de l'équivalence qui existe entre une construction-of/by et une construction avec possessif, on peut obtenir l'équivalence illustrée par le schéma 4.19.



#### 4.2.4 Les prépositions, marqueurs argumentaux du GNpréd

Si le rôle des arguments du GNpréd sont difficilement prédictibles quand ils sont placés à gauche du nom (possessifs ou prémodifieurs) leur position comme post-modifieurs rend leur rôle plus facilement déductibles dans la mesure où les différentes prépositions, introductrices de GN, peuvent être considérées comme marqueurs de ces arguments. Il faut tout d'abord noter qu'aucune préposition à elle seule n'a ce pouvoir de marquage. Chaque prédicat nominal sélectionne, en fonction de la catégorie verbale duquel il dérive morphologiquement, la ou les prépositions candidates qui marquent ces arguments. Ces prépositions font partie des entrées lexicales de ses nominalisations. Ainsi, il est possible de mettre en évidence deux grandes catégories de nominalisations : celles pour lesquelles la préposition *of* introduit un argument COD que l'on qualifie aussi de génitif objectif et celles pour lesquelles cette même préposition introduit un argument sujet nommé aussi génitif subjectif. Nous verrons enfin que génitif objectif et subjectif peuvent coexister dans un GNpréd.

##### La préposition *of* comme sélecteur du COD

Le fait qu'un prédicat nominal puisse être associé à un verbe à COD permet de décider que la préposition *of* sera le marqueur de ce COD dans le GNpréd. Rentre dans cette catégorie les nominalisations dont les verbes associés appartiennent au schéma verbal suivant :  $N_0 V N_1$  [Prep  $N_2 \dots$  Prep  $N_n$ ]. Ici et dans la suite, les crochets ("[" et "]") signifient que les éléments placés entre peuvent être effacés. Les verbes avec COD, ayant, pour la plupart d'entre eux, un emploi passif, la préposition *by* marquera de la même façon, dans cet emploi verbal et dans le GNpréd, l'argument sujet. L'exemple 4.20a et 4.20b montre une phrase dans son emploi actif et passif. L'équivalent nominal de ces phrases (exemple 4.20c) montre que la préposition *by* du passif marque un groupe nominal sujet dans le GNpréd (ici "*the surgeon*"). La préposition *of* qui n'apparaît que dans le GNpréd est le marqueur du COD ("*the patient*"). L'exemple 4.21a n'est qu'une variante du précédent. Les prépositions *of* et *by* introduisent également le COD ("*the painting*") et le sujet ("*the expert*"). Le second argument prépositionnel ("*Picasso*") introduit dans la phrase par la préposition *to* se retrouve dans le GNpréd introduit par la même préposition (exemple 4.21b).

(4.20) (a) *the surgeon amputated the [patient]<sub>Cod</sub>*

(le chirurgien a amputé le patient)

(b) *the patient was amputated by the [surgeon]<sub>Sujet</sub>*

(le patient a été amputé par le chirurgien)

(c) *the amputation of the [patient]<sub>Cod</sub> by the surgeon*

(l'amputation du patient par le chirurgien)

(4.21) (a) *the expert attributed the [painting] to Picasso*

(les experts ont attribué le tableau à Picasso)

(b) *the attribution of the painting to Picasso by the expert*

(l'attribution du tableau à Picasso par les experts)

Le schéma argumental de ce type de GNpréd est donc le suivant :

$N_v$  [of  $N_1$ ] [by  $N_0$ ] [Prep  $N_2$  . . . Prep  $N_n$ ]

### La préposition *of* comme sélecteur du sujet

Les GNpréd associés à une forme verbale sans COD, présentent la particularité d'introduire le sujet dans le GNpréd avec la préposition "of". Les verbes associés à ce type de nominalisation correspondent à un schéma verbal sans complément ou avec une suite quelconque de compléments prépositionnels que l'on peut noter :  $N_0$  V [Prep  $N_1$  . . . Prep  $N_n$ ]. Ces compléments, s'ils existent, se retrouvent dans le GNpréd, introduits par les mêmes prépositions que celles de la forme verbale. Ainsi l'exemple 4.22a, qui correspond à un emploi verbal sans complément, "the cow", l'argument sujet est introduit avec la préposition *of* (exemple 4.22b). De la même façon, pour l'exemple 4.23a, l'argument sujet "the pilot", est introduit comme prévu par *of*. L'argument complément ("the flight plan route") se retrouve dans le GNpréd introduit par la même préposition que pour la forme verbale "from" (exemple 4.23b).

(4.22) (a) *the cow calved*

(la vache a vêlé)

(b) *the calving of the [cow]<sub>sujet</sub>*

(le vêlage de la vache)

(4.23) (a) *the [pilote]<sub>Sujet</sub> deviated from the flight plan route*

(le plan a dévié du plan de vol)

(b) *the deviation of the pilot from the flight plan route*

(la déviation du pilote du plan de vol)

On aura, pour ce type de GNpréd le schéma argumental suivant :  
 $N_v$  [of  $N_0$ ] [Prep  $N_1$  . . . Prep  $N_n$ ].

### Les doubles constructions en *of*

Les GNpréd avec un double génitif (objectif et subjectif) sont plus rares. D'une façon générale, ils ne dénotent pas un processus mais une variante aspectuelle d'un processus qui est arrivé à son terme, que l'on peut considérer comme le résultat de ce processus.

(4.24) (a) *Peter translated the work of Zola* [Alexiadou, 2001]

(Peter a traduit l'œuvre de Zola)

(b) *the translation of Peter of the work of Zola*

(la traduction de Peter de l'œuvre de Zola)

(4.25) (a) *the enemy destroyed the city* [Alexiadou, 2001]

(l'ennemi a détruit la cité)

(b) *the destruction of the city by/\*of the enemy*

(la destruction de la cité par/\*de l'ennemi)

L'exemple 4.24 montre que *translation* est le résultat de l'action, ce qui n'est pas le cas de *destruction* (exemple 4.25) qui est clairement relié à un processus. Pour cette raison le double génitif n'est possible qu'avec l'exemple 4.24. Ainsi que le souligne l'auteur, quand il s'agit de produire un GNpréd de type processus, le sujet ne peut apparaître sous la forme d'un génitif alors que celui-ci apparaîtra systématiquement sous la forme d'un PP de type *by-phrase*. Notons que dans la suite de ce travail nous ne traiterons pas les doubles génitifs.

Nous avons vu dans les sections précédentes que la structure distributionnelle d'un GNpréd se compose principalement d'un nom de tête prédicatif et des arguments. Ces derniers peuvent être dans une position gauche avant le prédicat ou en position droite après le prédicat. Dans le dernier cas, les arguments sont introduits par différentes prépositions. Dans la section 4.3, nous décrirons comment nous avons utilisé la plate-forme PredicateDB pour confirmer le rôle que peuvent jouer ces prépositions dans les différents GNPréd et pour construire un lexique de prédicats nominaux.

## 4.3 Une base de données pour l'acquisition des prédicats nominaux

Nous avons vu (*Cf.* Section 4.2.2) qu'il n'existe pas de marqueurs qui nous permettent de déterminer les fonctions syntaxiques des arguments nominaux lorsqu'ils sont en position gauche (déterminant et prémodifieur). Par contre, les prépositions peuvent être considérées comme des marqueurs fiables pour déterminer les fonctions syntaxiques. Dans le but de confirmer le rôle que jouent les prépositions dans les différents types de GNpréd, nous avons développé PredicateDB, une plate-forme organisée autour d'une base de données et des différents outils permettant sa manipulation. Les données ont été extraites du Specialist Lexicon (*Cf.* Section 2.2.3). Cette plate-forme nous permet de confirmer les liens qui existent entre les schémas verbaux et les GNpréd qui leur correspondent en exploitant les propriétés syntaxiques des prédicats verbaux et nominaux fournies par le SL. Cette confirmation passe par la connaissance du comportement syntaxique des arguments lorsqu'ils sont dans un schéma verbal et lorsqu'ils se retrouvent dans le schéma nominal correspondant. Cet outil nous a permis également d'extraire un ensemble de nominalisations en fonction de leur comportement syntaxique dans le but de construire semi-automatiquement un lexique de nominalisations. Ce dernier a été utilisé dans le développement d'une grammaire de liens qui a été intégrée dans la grammaire du Link Parser [Sleator and Temperley, 1991] dans le but de rendre possible l'analyse syntaxique des GNpréd dans le domaine biologique [Royauté et al., 2006, 2007].

### 4.3.1 L'architecture générale de PredicateDB

PredicateDB a été conçu de telle sorte que toutes les informations relatives aux verbes et à leurs nominalisations soient éclatées. Cela nous donne une grande flexibilité qui nous permet d'accéder spécifiquement à chaque propriété syntaxique verbale et nominale d'une façon individuelle. Grâce à son architecture, PredicateDB permet d'accéder à toutes les informations qui appartiennent à n'importe quel schéma verbal ainsi qu'aux compléments nominaux qui leur correspondent et de croiser ces informations dans le but de rendre l'extraction plus efficace. PredicateDB offre plus de possibilités que LexAccess (Lexical Access Tool), qui est un outil d'interrogation développé par les concepteurs du Specialist Lexicon (Lexical Systems Group<sup>2</sup>). LexAccess ne permet l'extraction que d'un nombre réduit d'informations à partir du SL : la forme de base ("base"), l'identificateur de l'entrée ("entry"), la catégorie syntaxique ("cat") et les entrées qui appartiennent à une catégorie syntaxique donnée.

Nous avons choisi d'utiliser le Specialist Lexicon plutôt que d'autres ressources électroniques

---

2. <http://lexsrv3.nlm.nih.gov/LexSysGroup>

(NOMLEX-PLUS (Cf. Section 2.2.2), WordNet (Cf. Section 2.3.1), FrameNet (Cf. Section 2.3.2), etc.) pour les raisons suivantes :

- il contient une importante nomenclature de termes qui appartiennent principalement au domaine biologique (noms, verbes, adjectifs, etc.)
- il fait le lien entre les verbes et leurs nominalisations. Ainsi, nous pouvons connaître la ou les nominalisation(s) associée(s) à un verbe donné et vice-versa.
- il décrit les schémas verbaux de chaque verbe en spécifiant : (i) la catégorie syntaxique du verbe (transitif direct, intransitif, etc.), (ii) les prépositions lorsqu'elles sont présentes dans le schéma verbal et le type des compléments qu'elles introduisent, (iii) des propriétés syntaxiques comme : la possibilité ou non d'admettre une construction passive, les différents contrôles appliqués sur les arguments (contrôle sur l'objet<sup>3</sup>, contrôle sur le sujet<sup>4</sup>, etc.)
- il décrit partiellement les schémas nominaux en fournissant certaines informations : (i) les prépositions qui introduisent les compléments de nom, (ii) le type des compléments introduits par ces prépositions (syntagmes nominaux, gérondives, complétives, etc.), (iii) les différents contrôles qui peuvent s'appliquer sur les arguments nominaux, etc.

Nous avons déjà vu dans le chapitre 3.4.3 comment était conçu le SL. Nous rappelons ici sa structure à partir du verbe *dictate* (dicter) et sa nominalisation *dictation* dans le SL (Cf. Figure 4.3). Le verbe *dictate* admet les schémas verbaux *intran*, *tran=np* et *ditran=np,pphr(to,np)*. Ces derniers signifient que le verbe *dictate* peut être dans un schéma intransitif ( $N_0 V$ ), transitif direct ( $N_0 V N_1$ ) et dans un schéma ditransitif direct qui admet un COD et un complément prépositionnel composé d'un syntagme nominal (*np*) introduit par la préposition *to* ( $N_0 V N_1 to N_2$ ). La figure montre aussi que la nominalisation *dictation* admet des compléments prépositionnels (*pphr*) qui se composent de syntagmes nominaux (*np*) introduits par les prépositions *to*, *of* et *by*. PredicateDB traite 3 760 verbes et 3 960 nominalisations, ce qui nous a permis d'accéder aux propriétés syntaxiques de 905 schémas verbaux et 121 schémas nominaux, soit au total 1 026 schémas<sup>5</sup>.

3. Un contrôle sur l'objet signifie que le COD de la proposition principale (du verbe principal) est en même temps le sujet de la subordonnée enchassée. Par exemple, le COD *Marguerite* de la phrase *I advised Marguerite to resolve the problem* (J'ai conseillé Marguerite de résoudre le problème) est aussi le sujet de la proposition enchassée *to resolve the problem*

4. Un contrôle sur le sujet signifie que le sujet de la phrase principale est aussi le sujet de la proposition subordonnée. Par exemple, le sujet *Michael* de la phrase *Michael promised to repair the car* (Michael a promis de réparer la voiture) est aussi le sujet de la subordonnée infinitive *to repair the car*

5. Deux schémas qui possèdent le même type syntaxique, par exemple transitif indirect ( $N_0 V prep N_1$ ) mais dont certaines propriétés syntaxiques (les prépositions, le type des compléments, etc.) ne sont pas les mêmes sont considérés comme des schémas différents.

```

{base=dictate
entry=E0022504
cat=verb
variants=reg
intran
tran=np
ditran=np,pphr(to,np)
nominalization=dictation|noun|E0022505
}

{base=dictation
entry=E0022505
cat=noun
variants=reg
variants=uncount
compl=pphr(to,np)
compl=pphr(of,np)
compl=pphr(by,np)
nominalization_of=dictate|verb|E0022504
}

```

FIGURE 4.3 – La représentation du verbe *dictate* et de sa nominalisation *dictation* dans le Specialist Lexicon

### 4.3.2 Les différents modules de PredicateDB

PredicateDB se compose de trois modules : (i) le module Extract and Split (EAS) qui extrait les données intéressantes à partir du Specialist Lexicon (SL), (ii) le module BD qui permet de créer la base de données à partir des données fournies par le lexique SL et (iii) le module Post-Processing for Lexical Creation (PPLC) qui a pour fonction d'effectuer du post-traitement sur les résultats des requêtes lancées sur la base de données.

1. **le module EAS (Extract And Split) :** Ce module se charge d'extraire à partir du Specialist Lexicon toutes les données qui concernent les verbes et leurs nominalisations. Il a pour rôle de décomposer tout schéma verbal ou nominal et d'en extraire toutes les propriétés qui les caractérisent. Ces données serviront à enrichir les différentes tables de la base de données. Par exemple, à partir de l'entrée *dictate* (Figure 4.3), on peut extraire les informations suivantes :
  - l'identificateur de l'entrée (*E0022504*)
  - la forme de base de l'entrée (*dictate*)

- la nominalisation associée à l’entrée (*dictation*)
  - l’identificateur de cette nominalisation (*E0022505*)
  - les différentes catégories syntaxiques du verbe *dictate* : intransitif, transitif direct et ditransitif direct
  - toutes les propriétés qui caractérisent les schémas de complémentation associées aux différentes catégories syntaxiques : la présence du COD pour les schémas transitif direct et ditransitif direct, le type du complément d’objet direct (syntagme nominal pour les deux schémas), la présence de la préposition *to* dans le schéma ditransitif direct et le type (syntagme nominal) du complément introduit par cette préposition.
2. **le module DB (Data Base)** : il s’agit d’une base de données Oracle dont les tables ont été créées à partir des données extraites précédemment avec le module EAS. À partir de ces données, on a pu créer cinq tables. Ces dernières sont représentées par la figure 4.4.
- la table 1 associe chaque verbe ou nominalisation à son identificateur d’entrée unique. Par exemple, le verbe *dictate* est associé à l’id-entry *E0022504*. Cette dernière représente la clé primaire. La table contient 12 545 entrées.
  - la table 2 nous permet de faire correspondre chaque verbe à sa ou ses nominalisations. Par exemple, l’identificateur du verbe *dictate* (*E0022504*) est associé à celui de la nominalisation *dictation* (*E0022505*). La table contient 3 993 entrées et la clé est représentée par le couple (*id-nominalisation*, *id-verbe*).
  - la table 3 associe chaque *id-entry* d’un verbe ou d’une nominalisation à un ou plusieurs *id-schema*. Ce dernier fait référence à la classe grammaticale du schéma ainsi qu’aux propriétés syntaxiques de son schéma de complémentation. Par exemple, le verbe *dictate* est associé à trois *id-schemas* : à l’*id-schema* 005 qui fait référence au schéma intransitif (*intran*), à l’*id-schema* 001 qui fait référence au schéma transitif direct (*tran=np*) et à l’*id-schema* 002 qui fait référence au schéma ditransitif direct avec un COD et *to* comme introducteur du syntagme nominal *np* (*ditran=np,pphr(to,np)*). La nominalisation *dictation* est associée à l’*id-schema* 003 qui fait référence à un schéma dans lequel le complément de nom *np* est introduit par la préposition *of* (*compl=pphr(of,np)*), à l’*id-schema* 004 qui fait référence à un complément de nom introduit par la préposition *by* (*compl=pphr(by,np)*) et à l’*id-schema* 025 qui fait référence à un schéma nominal dans lequel le complément de nom est introduit par la préposition *to*

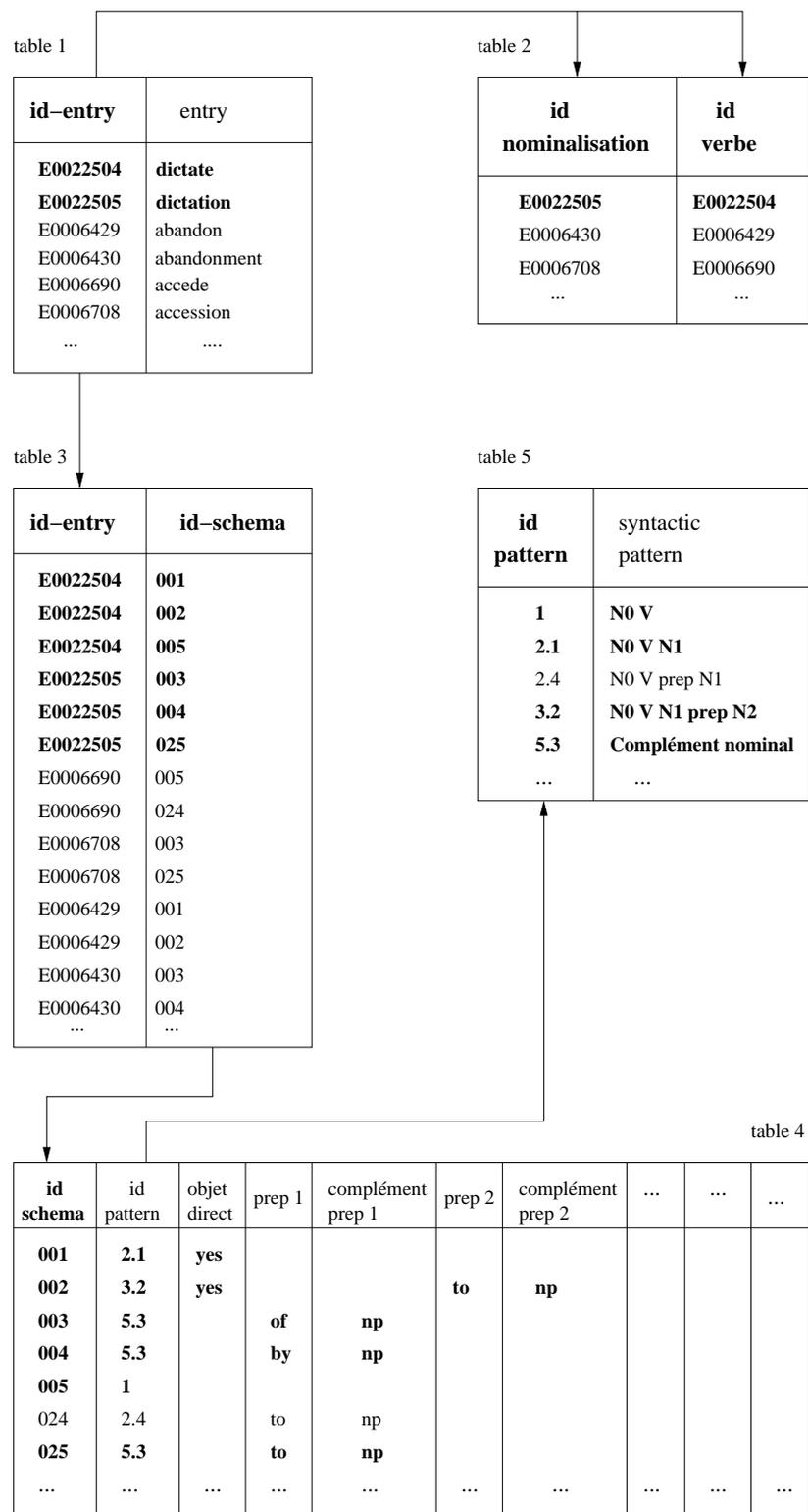


FIGURE 4.4 – La structure de la base de données

(`compl=pphr(to,np)`). Deux schémas verbaux ont des id-schemas différents s'ils ne possèdent pas la même classe grammaticale ou s'ils n'ont pas les mêmes propriétés syntaxiques. Par exemple, les deux schémas verbaux `tran=pphr(to,np)` et `tran=pphr(at,np)` possèdent la même classe grammaticale (transitif indirect) mais ne possèdent pas les mêmes propriétés syntaxiques car ils n'ont pas la même préposition. Ces deux schémas auront des id-schemas différents. De même, les deux schémas verbaux `tran=pphr(in,np)` et `tran=np;nopass` n'ont pas la même id-schema car ils ne possèdent pas la même classe grammaticale. Le premier schéma est transitif indirect et le deuxième est transitif direct. On traite les schémas nominaux de la même manière. Deux schémas nominaux possèdent des id-schemas identiques s'ils ont les mêmes propriétés syntaxiques. Par exemple, les schémas de complémentation nominaux `compl=pphr(of,np)` et `compl=pphr(by,np)` de la nominalisation *dictation* ont des id-schemas différents car ils diffèrent au niveau des prépositions qu'ils admettent. On a besoin de la clé (`id-entry`, `id-schema`) pour pouvoir accéder aux 27 659 entrées de la table.

- la table 4 est le résultat de la décomposition des schémas verbaux et nominaux. À l'aide du module EAS, on fait éclater chaque schéma verbal ou nominal, extrait toutes les propriétés syntaxiques qui le composent et on les met dans cette table. Cette dernière contient 29 colonnes, chaque colonne représente une propriété syntaxique particulière. Dans le schéma de la figure 4.4 nous n'avons représenté que les colonnes qui permettent d'illustrer notre exemple. On accède à chaque entrée appartenant à cette table en utilisant la clé primaire `id-schema`. Pour pouvoir extraire les schémas en ne se basant que sur la classe grammaticale des schémas, on associe à chaque entrée un identificateur de pattern (`id-pattern`). Ce dernier fait référence uniquement à la classe grammaticale du schéma et ne tient pas compte de la différence qui peut exister au niveau des propriétés syntaxiques, c'est-à-dire qu'il ne tient pas compte des différences qui existent au niveau : des prépositions, du nombre des compléments, la permission ou non du passif, les différents contrôles, etc. Par exemple, Les trois schémas qu'admet le verbe *dictate* (`intran`, `tran=np` et `ditran=np,pphr(to,np)`) possèdent trois `id-pattern` différents car ils diffèrent au niveau de la classe grammaticale. On associe au premier schéma l'`id-pattern` 1 qui veut dire que le schéma est intransitif. Le deuxième schéma possède l'`id-pattern` 2.1 où le chiffre 2 signifie que le schéma est transitif et le chiffre 1 qu'il possède un complément d'objet direct. Le troisième schéma possède l'`id-pattern` 3.2 où le chiffre 3 veut dire qu'il est ditransitif et le chiffre 2 qu'il

est direct. C'est-à-dire qu'il accepte deux compléments : le premier est un COD et le deuxième est un complément prépositionnel. La préposition *to* qui introduit ce dernier est mise dans une autre colonne dont l'étiquette est `Prep`. Le type du complément introduit par la préposition *to* est spécifié dans la colonne `complément prep` 2. Il est considéré dans ce cas comme un syntagme nominal (`np`). Les schémas nominaux associés à la nominalisation *dictation* possèdent le même `id-pattern` (5.3). Le chiffre 5 signifie qu'il s'agit d'un schéma nominal et le chiffre 3 veut dire que ce schéma admet un complément de nom introduit par une préposition. Cette table contient 1 026 entrées.

- la table 5 nous permet d'associer à chaque `id-pattern` un schéma syntaxique générique. Un schéma générique écrit avec la notation de Maurice Gross (Cf. Section 4.2.2) et dans lequel on ne spécifie que la catégorie syntaxique (intransitif, transitif direct, etc.) et le type des compléments admis (complément de nom, complément prépositionnel, etc.). Les autres propriétés syntaxiques (les prépositions, type des contrôles, etc.) ne sont pas prises en considération. Cette table a pour rôle d'associer à chaque schéma verbal ou nominal une notation claire et compréhensive pour l'utilisateur qui décrit le schéma syntaxique. Par exemple, on associe aux `id-pattern` 1, 2.1 et 3.2 du verbe *dictate* les schémas génériques :  $\langle N_0 V \rangle$  (`intran`),  $\langle N_0 V N_1 \rangle$  (`tran=np`) et  $\langle N_0 V N_1 Prep N_2 \rangle$  (`ditran=np,pphr(pre, np)`). Cette table décrit 47 schémas génériques.

3. **le module PPLC (Post-processing for Lexical Creation)** : Ce module est un ensemble de programmes Perl qui prennent en entrée les données qui sont le résultat de l'interrogation de la base de données. Il effectue un post-traitement afin de faciliter leur manipulation dans le but de pouvoir créer notre lexique de nominalisations.

### 4.3.3 Détermination des types argumentaux

La construction et l'utilisation de la base de données reposent sur des informations linguistiques connues, bien que rarement formalisées, sur les relations existant entre une structure syntaxique verbale et le groupe nominal associé (dont le nom de tête est la nominalisation de ce verbe) qui véhicule la même information. Une typologie de ces principales relations est donnée dans la table 4.1. Cette typologie exclue les compléments phrastiques et porte uniquement sur les compléments nominaux.

Sur la base de cette typologie, PredicateDB facilite l'étude des schémas verbaux et nominaux

qui appartiennent au SL en nous permettant de traiter un grand nombre de schémas. Il nous évite ainsi un parcours uniquement manuel des schémas verbaux et les schémas de complémentation des nominalisations qui leur sont associées et permet de privilégier un parcours intelligent fondé sur des hypothèses. L'interrogation de la base de données et l'utilisation des programmes Perl nous permettent d'extraire toutes les données qui sont intéressantes. En extrayant les propriétés syntaxiques des schémas verbaux et en les croisant avec les informations qui caractérisent les schémas nominaux correspondant, nous pouvons confirmer les rôles que jouent les différentes prépositions et spécialement la préposition *of*. Ceci nous permet de compléter les schémas de complémentation des différentes nominalisations qui sont par défaut incomplets dans le SL. Notre étude ne concerne pas tous les types des schémas verbaux présents dans le SL mais uniquement les schémas représentés par la table 4.1. Les schémas verbaux traités sont les suivants :

- un schéma intransitif ( $N_0 V$ ) dans lequel le seul argument est le sujet ;
- un schéma transitif direct ( $N_0 V N_1$ ) avec un sujet et un COD ;
- un schéma transitif indirect ( $N_0 V \text{ prep } N_1$ ). Ce schéma contient un sujet et un complément prépositionnel introduit par une préposition ;
- un schéma ditransitif direct ( $N_0 V N_1 \text{ prep } N_2$ ) qui admet un sujet, un COD et un complément prépositionnel introduit par une préposition ;
- un schéma ditransitif indirect ( $N_0 V \text{ prep}_1 N_1 \text{ prep}_2 N_2$ ) dans lequel on trouve un sujet et deux compléments prépositionnels introduits par deux prépositions différentes ;
- un schéma ditransitif ( $N_0 V \text{ prep } N_2 N_1$ ) qui possède un COD ( $N_1$ ) positionné après le groupe prépositionnel  $\langle \text{prep } N_2 \rangle$ . On peut rencontrer ce type de schéma lorsque le COD est trop long par rapport au complément prépositionnel ;
- un schéma ditransitif ( $N_0 V N_2 N_1$ ) composé d'un COD ( $N_1$ ) placé après un complément d'objet indirect ( $N_2$ ).

Afin d'illustrer l'utilisation de PredicateDB pour confirmer les rôles des différentes prépositions et compléter les schémas de complémentations nominaux, prenons par exemple, le schéma verbal  $\langle N_0 V \rangle$  de la table 4.1. À l'aide d'une requête SQL, nous avons extrait les informations suivantes : tous les verbes qui n'admettent que le schéma intransitif direct, les nominalisations qui leur sont associées ainsi que leurs compléments nominaux. Cette requête a retourné 195 verbes et 200 nominalisations. En analysant les propriétés syntaxiques des différents compléments de nom des nominalisations, nous avons remarqué qu'ils admettent, dans la plupart des cas, la préposition *of*. Comme le schéma intransitif direct n'admet qu'un seul argument qui est le sujet (Cf. Section 4.2.4), il est possible de confirmer que la préposition *of*

<b>Schemas</b>	<b>Exemples</b>	<b>Les GNpréd correspondant</b>
1. N <sub>0</sub> V intran	<i>Marry disappeared</i> (Marry a disparu)	N <sub>V</sub> of N <sub>0</sub> <i>The disappearance of Marry</i> (la disparition de Marry)
2. N <sub>0</sub> V N <sub>1</sub> tran=np	<i>Canada achieves reforms</i> (le Canada a réalisé des réformes)	N <sub>V</sub> of N <sub>1</sub> by N <sub>0</sub> <i>the achievement of reforms by Canada</i> (la réalisation des réformes par le Canada)
3. N <sub>0</sub> V Prep N <sub>1</sub> tran=pphr	<i>the child arrived at home</i> (l'enfant est arrivé à la maison)	N <sub>V</sub> of N <sub>0</sub> Prep N <sub>1</sub> <i>the arrival of the child at home</i> (l'arrivée de l'enfant à la maison)
4. N <sub>0</sub> V N <sub>1</sub> Prep N <sub>2</sub> ditran=np,pphr	<i>The experts attributed the painting to Picasso</i> (les experts ont attribué le tableau à Picasso)	N <sub>V</sub> of N <sub>1</sub> Prep N <sub>2</sub> by N <sub>0</sub> <i>The attribution of the painting to Picasso by the experts</i> (l'attribution du tableau à Picasso par les experts)
5. N <sub>0</sub> V Prep <sub>1</sub> N <sub>1</sub> Prep <sub>2</sub> N <sub>2</sub> ditran=pphr,pphr	<i>the employee converses on the Internet with his boss</i> (l'employé converse sur internet avec son patron)	N <sub>V</sub> of N <sub>0</sub> Prep <sub>1</sub> N <sub>1</sub> Prep <sub>2</sub> N <sub>2</sub> <i>the conversation of the employee on the Internet with his boss</i> (la conversation de l'employé sur internet avec son patron)
6. N <sub>0</sub> V Prep N <sub>2</sub> N <sub>1</sub> ditran=pphr,np	<i>U.S explained to Canada the fifth protocol</i> (les États-Unis ont expliqué au Canada le cinquième protocole)	N <sub>V</sub> of N <sub>1</sub> Prep N <sub>2</sub> by N <sub>0</sub> <i>the explanation of the fifth protocol to Canada by U.S</i>  (l'explication du cinquième protocole au Canada par les États-Unis)
7. N <sub>0</sub> V N <sub>2</sub> N <sub>1</sub> ditran=np,np	<i>the chef prepared the guest breakfast</i> (le chef a préparé un petit-déjeuner aux invités)	N <sub>V</sub> of N <sub>1</sub> Prep N <sub>2</sub> by N <sub>0</sub> <i>the preparation of breakfast for the guest by the chef</i> (la préparation d'un petit-déjeuner aux invités par le chef)

TABLE 4.1 – Les relations entre certains schémas verbaux et les GNpréd correspondant

est dans ce cas le marqueur privilégié du sujet dans le GNpréd associé.

Cette démarche nous a permis aussi d'observer certains faits linguistiques, comme l'existence d'un lien entre les schémas  $\langle N_0 V \rangle$ ,  $\langle N_0 V \text{ prep } N_1 \rangle$  et  $\langle N_0 V \text{ prep}_1 N_1 \text{ prep}_2 N_2 \rangle$ . En effet, on a observé que la plupart des verbes qui admettent le schéma  $\langle N_0 V \text{ prep}_1 N_1 \text{ prep}_2 N_2 \rangle$ , admettent en plus, soit le schéma  $\langle N_0 V \text{ prep } N_1 \rangle$ , soit le schéma  $\langle N_0 V \rangle$  soit les deux. L'exemple 4.26a, qui est associé au schéma  $\langle N_0 V [\text{prep}_1 N_1] [\text{prep}_2 N_2] \rangle$  (les arguments entre crochets sont optionnels), montre qu'on peut passer du schéma  $\langle N_0 V \text{ prep}_1 N_1 \text{ prep}_2 N_2 \rangle$  au schéma  $\langle N_0 V \text{ prep } N_1 \rangle$  en effaçant *to China*. Si on efface encore *from Canada*, la phrase résultante est associée au schéma  $\langle N_0 V \rangle$ . Les GNpréd associés à ces trois schémas verbaux, qui sont ici représentés par le GNpréd 4.26b, montrent qu'ils admettent la préposition *of* comme marqueur du sujet. Ceci nous permet de dire que ces schémas possèdent des propriétés syntaxiques communes.

- (4.26) (a) *Joanna remigrated [from Canada] [to China]*  
 (Joanna a réémigré [du Canada] [vers la Chine])  
 (b) *the remigration of Joanna<sub>sujet</sub> [from Canada] [to China]*  
 (la réémigration de Joanna [du Canada] [vers la Chine])

D'une manière générale, le lien qui relie les schémas verbaux sans COD aux GNpréd qui leur correspondent est représenté par le schéma 4.27 où la préposition *of* est, dans la plupart des cas, le marqueur du sujet. Les arguments entre crochets sont optionnels. Par exemple, à partir du schéma  $\langle N_0 V \text{ prep}_1 N_1 \text{ prep}_2 N_2 \rangle$ , on peut obtenir le schéma  $\langle N_0 V \text{ prep}_1 N_1 \rangle$  en effaçant le groupe prépositionnel  $\langle \text{prep}_2 N_2 \rangle$ .

$$N_0 V [\text{prep}_1 N_1] [\text{prep}_2 N_2] \rightleftharpoons N_v \text{ of } N_0 [\text{prep}_1 N_1] [\text{prep}_2 N_2] \quad (4.27)$$

Nous adoptons la même méthode pour confirmer que les prépositions *of* et *by* sont respectivement les marqueurs du complément d'objet direct et du sujet dans le schéma transitif direct ( $N_0 V N_1$ ). En extrayant à partir de la base de données tous les verbes qui n'admettent que le schéma transitif direct et leurs nominalisations - cette requête a permis de retourner 1 812 verbes et 1 839 nominalisations - nous observons que la plupart des compléments de nom de ces nominalisations se construisent avec les prépositions *of* et *by*. Ces informations nous ont permis de confirmer que le *by* est le marqueur du sujet et que le *of* est le marqueur du COD (Cf. Section 4.2.4). À l'instar du lien qui existe entre les schémas intransitifs, nous avons observé l'existence d'un lien entre certains schémas transitifs. En extrayant tous les verbes

qui admettent un schéma de type  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  et leurs nominalisations - la requête a renvoyé 676 verbes et 729 nominalisations - nous avons remarqué que la plupart des GNpréd admettent les prépositions *of* et *by* et que la majorité des verbes qui admettent ce type de schéma, admettent également le schéma  $\langle N_0 V N_1 \rangle$  (exemple 4.28a). Ces deux informations confirment les données du tableau 4.1 en montrant qu'il existe un lien syntaxique entre les schémas  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  et  $\langle N_0 V N_1 \rangle$ . Ceci confirme également que le GNpréd associé à un schéma de type  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  admettent les prépositions *of* et *by* comme marqueurs du COD et du sujet. Il est intéressant de noter aussi que les compléments prépositionnels présents dans les schémas verbaux sont conservés dans les GNpréd correspondants et qu'ils sont introduits par les mêmes prépositions. Les exemples 4.28a 4.28b montrent qu'on peut passer du schéma  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  au schéma  $\langle N_0 V N_1 \rangle$  en effaçant le complément *to AC*.

- (4.28) (a) *the microsomal fraction*<sub>Sujet</sub> *also biotransforms EtOH*<sub>Cod</sub> [*to AC*]  
 (la fraction microsomale biotransforme également l'EtOH [en AC])  
 (b) *the biotransformation of the ETOH* [*to AC*] *by microsomal fraction*  
 (la biotransformation de l'ETOH en AC par la fraction microsomale)

Le schéma 4.29 montre le lien qui existe entre les schémas verbaux qui admettent un COD ( $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  et  $\langle N_0 V N_1 \rangle$ ) et leurs GNpréd. Dans ce type de schémas, la préposition *of* est le marqueur du COD et le *by* est le marqueur du sujet.

$$N_0 V N_1 [\text{prep } N_2] \rightleftharpoons N_v \text{ of } N_1 [\text{prep } N_2] \text{ by } N_0 \quad (4.29)$$

Nous avons aussi pu confirmer le GNpréd associé au schéma  $\langle N_0 V \text{ prep } N_2 N_1 \rangle$ . Il est caractérisé par le fait que son COD est placé après le groupe prépositionnel (exemple 4.30a). Ce schéma peut être considéré comme l'équivalent du schéma  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$ , comme le montre l'exemple 4.30b. En analysant les GNpréd associés à notre schéma, on remarque la présence des prépositions *of* et *by* qui sont également, dans la plupart des cas, les marqueurs du COD et du sujet. Ceci peut être confirmé par le fait que la majorité des verbes qui admettent ce schéma, admettent aussi le schéma  $\langle N_0 V N_1 \rangle$  ainsi que le schéma  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$ . La préposition qui introduit le complément prépositionnel dans le schéma verbal est aussi présente dans le GNpréd associé à la nominalisation (exemple 4.30b).

- (4.30) (a) *the Republican Mitch McConnell appointed to the Commission Mr. Dennis C. Shea*

- (le Républicain Mitch McConnell a nommé à la tête de la commission M. Dennis C. Shea)
- (b) *the Republican Mitch McConnell appointed Mr. Dennis C. Shea to the Commission*  
(le Républicain Mitch McConnell a nommé M. Dennis C. Shea à la tête de la commission)
- (c) *the appointment to the Commission of Mr. Dennis C. Shea by the Republican Mitch McConnell*  
(la nomination à la tête de la commission de M. Dennis C. Shea par le Républicain Mitch McConnell)

Nous avons également confirmé le schéma  $\langle N_0 V N_2 N_1 \rangle$ . Ce dernier se caractérise par la présence de deux compléments nominaux en distribution inverse avec élision de la préposition pour  $N_2$ , c'est-à-dire, un complément d'objet indirect  $N_2$  suivi du COD  $N_1$  (exemple 4.31a). Ce schéma est équivalent au schéma  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  lorsque ce dernier autorise l'alternance du datif (exemple 4.31b). Les GNpréd associés à ce schéma possèdent les mêmes propriétés que celles du schéma précédent (exemple 4.31b).

- (4.31) (a) *the chef prepared the guest breakfast* [Reeves et al., 1999]  
(le chef a préparé un petit déjeuner aux invités)
- (b) *the chef prepared breakfast for the guest*
- (c) *the preparation of breakfast for the guest by the chef*  
(la préparation d'un petit déjeuner aux invités par le chef)

Une fois que tous les GNpréd associés aux schémas verbaux de la table 4.1 ont été confirmés et construits, nous pouvons compléter les GNpréd de chaque nominalisation appartenant au SL et associer chaque GNpréd au schéma verbal qui lui correspond. Prenons par exemple, le verbe *dictate* et sa nominalisation *dictation* (Figure 4.3). Le verbe *dictate* admet trois schémas verbaux : intransitif  $\langle N_0 V \rangle$ , transitif direct  $\langle N_0 V N_1 \rangle$  et ditransitif direct  $\langle N_0 V N_1 \text{ to } N_2 \rangle$ . La nominalisation *dictation* aura donc trois GNpréd : le GNpréd  $\langle \text{dictation of } N_0 \rangle$  dans lequel la préposition *of* introduit le sujet (associé au schéma intransitif) et les GNpréd  $\langle \text{dictation of } N_1 \text{ by } N_0 \rangle$  (transitif direct) et  $\langle \text{dictation of } N_1 \text{ by } N_0 \text{ to } N_2 \rangle$  (ditransitif direct) dans lesquels *of* introduit le COD.

Il est intéressant de noter qu'il existe deux manières d'interpréter le fait qu'un verbe admette plusieurs schémas verbaux. Quand un verbe donné admet par exemple les schémas

intransitif ( $N_0 V$ ), transitif direct ( $N_0 V N_1$ ) et ditransitif direct ( $N_0 V N_1 \text{ prep } N_2$ ), on peut comprendre que l'emploi  $\langle N_0 V \rangle$  correspond à un emploi  $\langle N_0 V N_1 \rangle$  où  $N_1$  peut être effacé et qu'un emploi  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  est un emploi  $\langle N_0 V N_1 \rangle$  auquel on a rajouté un groupe prépositionnel sans que le sens du verbe change, ou on peut comprendre que le verbe possède un sens différent à chaque emploi. Même si notre travail ne gère pas ce genre de problèmes, on a pu observer qu'on pouvait rencontrer les deux cas, cela dépend du verbe. Par exemple, le verbe *dictate*, qu'il soit dans un emploi transitif direct ou ditransitif direct (exemple 4.32), possède le sens de *dicter une lettre*.

(4.32) *the teacher dictated some sentences [to his students]*  
 (le professeur a dicté quelques phrases [à ses écoliers])

## 4.4 La création d'un lexique de nominalisations

Après avoir associé à chaque schéma verbal le GNpréd qui lui correspond, PredicateDB nous a permis de classer les nominalisations selon leurs comportements syntaxiques et d'utiliser cette classification pour créer un lexique. Dans le lexique, on associe à chaque entrée de nominalisation un type qui décrit les structures syntaxiques des GNpréd qu'elle peut admettre (les fonctions syntaxiques des arguments admis et les marqueurs qui les introduisent). Toutes les nominalisations qui possèdent les mêmes propriétés syntaxiques vont être marquées comme appartenant au même ensemble. Le lexique ne traite que les nominalisations dont les GNpréd appartiennent au tableau 4.1.

### 4.4.1 Classification des nominalisations

Notre tâche consiste à regrouper toutes les nominalisations dont les GNpréd possèdent la même structure syntaxique. Deux GNpréd possèdent la même structure syntaxique si les marqueurs (prépositions) qu'ils admettent introduisent les mêmes types d'argument. Ceci signifie qu'on a besoin de connaître la structure des GNpréd associés à chaque nominalisation afin d'effectuer une classification, or le SL ne fournit pas cette structure. Notre méthode a donc consisté alors à se baser sur les schémas verbaux associés aux nominalisations pour connaître la structure des GNpréd correspondants. En effet, les verbes qui possèdent les mêmes schémas verbaux se caractérisent par le fait que leurs nominalisations admettent le plus souvent des GNpréd qui possèdent les mêmes structures syntaxiques (Cf. Section 4.3.3). Par exemple, les phrases 4.33a et 4.33c montrent que les verbes *adhere* et *acquiesce* admettent le même schéma

transitif indirect ( $N_0$  V prep  $N_1$ ). En analysant les GNpréd qui leur correspondent (exemples 4.33b et 4.33d), nous pouvons remarquer qu'ils possèdent le même schéma syntaxique ( $N_v$  of  $N_0$  prep  $N_1$ ). Leurs schémas syntaxiques saturés possèdent les mêmes arguments : un sujet et un complément prépositionnel.

- (4.33) (a) *the paste adheres to the skin*  
 (la pâte adhère à la peau)  
 (b) *the adherence of the paste to the skin*  
 (c) *the father acquiesced in the removal of children . . .*  
 (le père a consenti au déplacement des enfants . . .)  
 (d) *the acquiescence of the father in the removal of children . . .*

Une fois les nominalisations classées selon les schémas verbaux associés à leurs verbes, l'étape suivante consiste à associer à chaque classe créée les GNpréd qui lui correspondent. La dernière étape consiste à regrouper les classes dont les GNpréd possèdent des structures syntaxiques similaires ou proches en super-classes.

**1) Regroupement des schémas verbaux :** Cette étape consiste à extraire à partir de la base de données tous les verbes qui possèdent les mêmes schémas verbaux et de mettre les nominalisations qui leur correspondent dans les mêmes classes. Par exemple, la figure 4.5 montre que les verbes *agree* (accepter, être d'accord), *reflect* (réfléter, réfléchir) et *arbitrate* (arbitrer) possèdent les mêmes schémas verbaux :  $\langle N_0$  V  $\rangle$ ,  $\langle N_0$  V  $N_1$   $\rangle$  et  $\langle N_0$  V prep  $N_1$   $\rangle$ . Les nominalisations associées à ces verbes ainsi que les autres nominalisations - dont les verbes possèdent les mêmes schémas verbaux - sont mis dans la même classe. Après avoir généralisé cette démarche sur le reste des schémas verbaux, nous avons obtenu 31 classes disjointes (Figure 4.6), où chaque nominalisation ne peut appartenir qu'à une seule classe.

**2) Association des GNpréd** Dans cette étape, on parcourt les classes qui ont été précédemment créées. On associe à chaque schéma verbal le GNpréd qui lui correspond puis on regroupe les GNpréd compatibles. Prenons, par exemple, la classe 16 de la figure 4.6. Cette classe contient les nominalisations qui correspondent aux verbes de la figure 4.5 (*agree*, *reflect* et *arbitrate*) ainsi que toutes les autres nominalisations associées aux mêmes schémas verbaux  $\langle N_0$  V  $\rangle$ ,  $\langle N_0$  V prep  $N_1$   $\rangle$  et  $\langle N_0$  V  $N_1$   $\rangle$ . D'après le tableau 4.1, les GNpréd associés à ces schémas sont respectivement :  $\langle N_v$  of  $N_0$   $\rangle$ ,  $\langle N_v$  of  $N_0$  prep  $N_1$   $\rangle$  et  $\langle N_v$  of  $N_1$  by  $N_0$   $\rangle$ . En analysant ces GNpréd, on remarque qu'on peut regrouper les GNpréd selon le rôle de

<pre>{base=agree entry=E0039294 intran tran=np tran=ppher (on , np) tran=ppher (to , np) tran=ppher (about , np) tran=ppher (with , np) nominalization=agreement}</pre>	<pre>{base=reflect entry=E0052424 intran tran=np tran=ppher (upon , np) tran=ppher (on , np) nominalization=reflection}</pre>
<pre>base=arbitrate entry=E0010256 intran tran=np tran=ppher (between , np) ; nopass nominalization=arbitration</pre>	

FIGURE 4.5 – Les schémas verbaux admis par les verbes *agree*, *reflect* et *arbitrate*

la préposition *of*, c'est-à-dire, selon l'argument nominal qu'elle introduit. On a vu (Cf. Section 4.3.3) que les GNpréd peuvent être divisés en deux groupes distincts : (i) ceux qui sont associés aux schémas verbaux transitifs, qui admettent un COD et dans lesquels la préposition *of* est le marqueur du COD et (ii) ceux qui sont associés aux schémas verbaux intransitifs, qui n'admettent pas de COD et dont la préposition *of* est le marqueur du sujet. En se basant sur cette propriété :

1. on associe aux deux schémas verbaux  $\langle N_0 V \rangle$  et  $\langle N_0 V \text{ prep } N_1 \rangle$  le GNpréd  $\langle N_v \text{ of } N_0 [\text{prep } N_1] \rangle$  dans lequel  $\langle \text{prep } N_1 \rangle$  est optionnel. Lorsque  $\langle \text{prep } N_1 \rangle$  est effacé, le GNpréd résultant  $\langle N_v \text{ of } N_0 \rangle$  est associé au schéma verbal intransitif direct  $\langle N_0 V \rangle$  et lorsqu'il est présent, il est associé au schéma verbal transitif indirect  $\langle N_0 V \text{ prep } N_1 \rangle$ . Dans les deux cas, la préposition *of* est le marqueur du sujet. Par exemple, la phrase 4.34a et son GNpréd 4.34b montrent qu'on peut passer d'un schéma  $\langle N_0 V \text{ prep } N_1 \rangle$  à un schéma  $\langle N_0 V \rangle$  en effaçant le groupe prépositionnel *on the price*.

- (4.34) (a) *Monica agreed [on the price]*  
(Monica est d'accord [sur le prix])  
(b) *the agreement of Monica [on the price]*

### Super-classe 1

- 1 ('ditran=np,np','ditran=np,pphr','tran=np','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$
  - 2 ('ditran=np,np','tran=np','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$
  - 3 ('ditran=np,pphr','tran=np','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$
  - 4 ('ditran=pphr,np','tran=np','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$
  - 5 ('ditran=np,np','intran')  $\rightarrow N_V$  of  $N_1$  Prep  $N_2$  by  $N_0$ ,  $N_V$  of  $N_0$
  - 6 ('ditran=np,pphr','intran')  $\rightarrow N_V$  of  $N_1$  Prep  $N_2$  by  $N_0$ ,  $N_V$  of  $N_0$
- 

### Super-classe 2

- 7 ('ditran=np,np','ditran=np,pphr','tran=np','tran=pphr','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$  [Prep  $N_1$ ]
  - 8 ('ditran=np,pphr','ditran=pphr,np','tran=np','tran=pphr','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$  [Prep  $N_1$ ]
  - 9 ('ditran=np,pphr','tran=np','tran=pphr','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$  [Prep  $N_1$ ]
  - 10 ('ditran=np,np','tran=np','tran=pphr','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$  [Prep  $N_1$ ]
  - 11 ('ditran=np,pphr','tran=pphr','intran')  $\rightarrow N_V$  of  $N_1$  Prep  $N_2$  by  $N_0$ ,  $N_V$  of  $N_0$  [Prep  $N_1$ ]
  - 12 ('ditran=np,pphr','tran=pphr')  $\rightarrow N_V$  of  $N_1$  Prep  $N_2$  by  $N_0$ ,  $N_V$  of  $N_0$  Prep  $N_1$
  - 13 ('ditran=np,pphr','tran=np','tran=pphr')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$  Prep  $N_1$
  - 14 ('ditran=np,pphr','tran=np','ditran=pphr,pphr','tran=pphr','intran')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$ ,  $N_V$  of  $N_0$  [Prep<sub>1</sub>  $N_1$ ] [Prep<sub>2</sub>  $N_2$ ]
  - 15 ('ditran=np,pphr','ditran=pphr,pphr','tran=pphr','intran')  $\rightarrow N_V$  of  $N_1$  Prep  $N_2$  by  $N_0$ ,  $N_V$  of  $N_0$  [Prep<sub>1</sub>  $N_1$ ] [Prep<sub>2</sub>  $N_2$ ]
- 

### Super-classe 3

- 16 ('tran=np','tran=pphr','intran')  $\rightarrow N_V$  of  $N_1$  by  $N_0$ ,  $N_V$  of  $N_0$  [Prep  $N_1$ ]
  - 17 ('tran=np','tran=pphr')  $\rightarrow N_V$  of  $N_1$  by  $N_0$ ,  $N_V$  of  $N_0$  Prep  $N_1$
  - 18 ('tran=np','intran')  $\rightarrow N_V$  of  $N_1$  by  $N_0$ ,  $N_V$  of  $N_0$
  - 19 ('tran=np','ditran=pphr,pphr','tran=pphr')  $\rightarrow N_V$  of  $N_1$  by  $N_0$ ,  $N_V$  of  $N_0$  Prep<sub>1</sub>  $N_1$  [Prep<sub>2</sub>  $N_2$ ]
- 

### Super-classe 4

- 20 ('ditran=np,np','ditran=np,pphr','tran=np')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$
  - 21 ('ditran=np,np','tran=np')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$
  - 22 ('ditran=np,pphr')  $\rightarrow N_V$  of  $N_1$  Prep  $N_2$  by  $N_0$
  - 23 ('ditran=np,pphr','ditran=pphr,np','tran=np')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$
  - 24 ('ditran=np,pphr','tran=np')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$
  - 25 ('ditran=pphr,np','tran=np')  $\rightarrow N_V$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$
  - 26 ('tran=np')  $\rightarrow N_V$  of  $N_1$  by  $N_0$
- 

### Super-classe 5

- 27 ('ditran=pphr,pphr','tran=pphr','intran')  $\rightarrow N_V$  of  $N_0$  [Prep<sub>1</sub>  $N_1$ ] [Prep<sub>2</sub>  $N_2$ ]
  - 28 ('ditran=pphr,pphr','tran=pphr')  $\rightarrow N_V$  of  $N_0$  Prep<sub>1</sub>  $N_1$  [Prep<sub>2</sub>  $N_2$ ]
- 

### Super-classe 6

- 29 ('tran=pphr','intran')  $\rightarrow N_V$  of  $N_0$  [Prep  $N_1$ ]
  - 30 ('tran=pphr')  $\rightarrow N_V$  of  $N_0$  Prep  $N_1$
  - 31 ('intran')  $\rightarrow N_V$  of  $N_0$
- 

FIGURE 4.6 – Les classes et les super-classes générées

(l'accord de Monica sur le prix)

2. on associe au schéma verbal transitif direct  $\langle N_0 \text{ V } N_1 \rangle$  le GNpréd  $\langle N_v \text{ of } N_1 \text{ by } N_0 \rangle$  dans lequel la préposition *of* introduit le COD et le *by* introduit le sujet (exemples 4.35a et 4.35b).

- (4.35) – *Monica agreed the deal*  
(Monica a accepté le marché)  
– *the agreement of the deal by Monica*  
(l'acceptation du marché par Monica)

L'association des GNpréd à leurs classes correspondantes est produite par l'algorithme 1. Le principe de cet algorithme est de parcourir l'ensemble des classes verbales et de leur associer les GNpréd qui leur correspondent. Cet algorithme prend en compte le fait que la plupart de ces classes verbales contiennent différents schémas verbaux. Dans cet algorithme, `tab-classesN[31]` représente un tableau d'enregistrements dans lequel sont définies les propriétés nominales de chaque classe en plusieurs champs (lignes 2 à 4) : le nombre de GNpréd (`nbreGnp`), le macrotype du groupe nominal `gnp0` quand la préposition *of* introduit le sujet et `gnp1` quand cette préposition introduit l'objet ainsi que le champ `traite` qui joue le rôle de marqueur et dont le rôle est détaillé dans l'algorithme 2. La fonction `obtenir-classe` (ligne 6) a pour rôle de classer les verbes selon les schémas verbaux qu'ils admettent et regroupe chaque ensemble de schémas dans une classe donnée. Cette fonction renvoie une table de hachage dont la clé est le numéro de la classe (`classe1`, `classe2`, ..., `classe31`) et les valeurs correspondent aux différents types de schémas. Le résultat de la fonction est affecté à la variable `hash-classesV` qui est de type table de hachage (ligne 6). Pour chaque élément `classei` appartenant au tableau `tab-classesN[ ]` (lignes 7 à 40), on effectue les opérations suivantes :

- on affecte à chaque classe<sub>i</sub> les deux GNpréd saturés  $\langle N_v \text{ of } N_1 \text{ prep } N_2 \text{ by } N_0 \rangle$  et  $\langle N_v \text{ of } N_0 \text{ prep } N_1 \text{ prep } N_2 \rangle$ . Cela est justifié par le fait que chaque classe<sub>i</sub> ne peut contenir que deux types de schémas verbaux : schémas verbaux transitifs avec COD et schémas verbaux intransitifs sans COD (lignes 8 et 9).
- dans la première étape (ligne 10 à ligne 19), on vérifie si la classe<sub>i</sub> contient un ou plusieurs schémas avec COD et on lui affecte le GNpréd qui correspond à ces schémas :
  1. si la classe<sub>i</sub> n'admet aucun schéma avec COD (ligne 10), le GNpréd `gnp1` a pour

---

**Algorithme 1** : L'algorithme qui permet d'associer les GNpréd à chacune des 31 classes

---

**Entrée** : 31 classes verbales

**Sortie** : Des GNpréd associés à chaque classe verbale

**1 Début**

```
2   Structure classes { entier : nbreGnp,traité ← 0 ;
3   chaîne de caractères : gnp1, gnp0;
4   } tab-classesN[31];
5   hash-classesV : tableau de hachage;
6   hash-classesV ← obtenir-classes();
7   Pour  $i \leftarrow 0$  à  $\text{taille}(\text{hash-classesV})-1$  faire
8       tab-classesN[i].gnp1 ←  $N_V$  of  $N_1$  Prep  $N_2$  by  $N_0$ ;
9       tab-classesN[i].gnp0 ←  $N_V$  of  $N_0$  Prep  $N_1$  Prep  $N_2$ ;
10      /* traitement du groupe nominal prédicatif GNP1 */;
11      Si ('ditran=np,pphr' et 'ditran=np,np' et 'ditran=pphr,np' et 'tran=np')  $\notin$ 
12      hash-classesV{i} Alors
13          | tab-classesN[i].gnp1 ←  $\emptyset$ ;
14      Sinon
15          | tab-classesN[i].nbreGnp ← tab-classesN[i].nbreGnp+1;
16          | Si ('ditran=np,pphr' et 'ditran=np,np' et 'ditran=pphr,np')  $\notin$ 
17          | hash-classesV{i} Alors
18              | tab-classesN[i].gnp1 ←  $N_v$  of  $N_1$  by  $N_0$  ;
19          | Sinon Si 'tran=np'  $\in$  hash-classesV{i} Alors
20              | tab-classesN[i].gnp1 ←  $N_v$  of  $N_1$  [Prep  $N_2$ ] by  $N_0$  ;
21          | FinSi
22      FinSi
23      /* traitement du groupe nominal prédicatif gnp0 */;
24      Si ('intran' et 'tran=pphr' et 'ditran=pphr,pphr')  $\notin$  hash-classesV{i} Alors
25          | tab-classesN[i].gnp0 ←  $\emptyset$  ;
26      Sinon
27          | tab-classesN[i].nbreGnp ← tab-classesN[i].nbreGnp+1;
28          | Si ('ditran=pphr,pphr')  $\notin$  hash-classesV{i} Alors
29              | tab-classesN[i].gnp0 ←  $N_V$  of  $N_0$  Prep  $N_1$ ;
30              | Si ('tran=pphr'  $\notin$  hash-classesV{i}) Alors
31                  | tab-classesN[i].gnp0 ←  $N_V$  of  $N_0$  ;
32              | Sinon Si ('intran'  $\in$  hash-classesV{i}) Alors
33                  | tab-classesN[i].gnp0 ←  $N_V$  of  $N_0$  [Prep  $N_1$ ] [Prep  $N_2$ ];
34              | Sinon Si ('tran=pphr'  $\in$  hash-classesV{i}) Alors
35                  | tab-classesN[i].gnp0 ←  $N_V$  of  $N_0$  Prep  $N_1$  [Prep  $N_2$ ];
36              | FinSi
37          | FinSi
38      FinSi
39      FinSi
40      FinPour
41 Fin
```

---

valeur l'élément vide (ligne 11).

2. le GNpréd  $gnp_1$  prend la valeur  $\langle N_v \text{ of } N_1 \text{ by } N_0 \rangle$  si la classe<sub>*i*</sub> n'admet aucun schéma ditransitif avec COD mais admet un schéma transitif direct (lignes 14 à 15).
  3.  $gnp_1$  prend la valeur  $\langle N_v \text{ of } N_1 \text{ [prep } N_2] \text{ by } N_0 \rangle$  si la même classe admet un schéma transitif direct en plus d'un ou plusieurs schémas ditransitifs avec COD (lignes 16 à 17).
- la deuxième étape consiste à vérifier si la classe<sub>*i*</sub> contient un ou plusieurs schémas intransitifs sans COD et on lui affecte les GNpréd correspondants (lignes 20 à 36) .
1. Si aucun des schémas  $\langle N_0 \text{ V} \rangle$ ,  $\langle N_0 \text{ V Prep } N_1 \rangle$  et  $\langle N_0 \text{ V Prep } N_1 \text{ Prep } N_2 \rangle$ , on affecte au GNpréd  $gnp_0$  l'élément vide (lignes 20 et 21).
  2. le GNpréd  $gnp_0$  prend la valeur  $\langle N_v \text{ of } N_0 \text{ prep } N_1 \rangle$  si la classe<sub>*i*</sub> contient les schémas précédents sauf le schéma  $\langle N_0 \text{ V prep } N_1 \text{ prep } N_2 \rangle$  (lignes 24 à 25).
  3.  $gnp_0$  prend la valeur  $\langle N_v \text{ of } N_0 \rangle$  si la classe ne contient ni le schéma  $\langle N_0 \text{ V prep } N_1 \text{ prep } N_2 \rangle$  ni le schéma  $\langle N_0 \text{ V Prep } N_1 \rangle$  mais le schéma  $\langle N_0 \text{ V} \rangle$  (lignes 26 à 27).
  4.  $gnp_0$  prend la valeur  $\langle N_v \text{ of } N_0 \text{ [prep } N_1] \rangle$  si la même classe contient le schéma  $\langle N_0 \text{ V Prep } N_1 \rangle$  ainsi que le schéma  $\langle N_0 \text{ V} \rangle$  (lignes 28 à 29).
  5.  $gnp_0$  prend la valeur  $\langle N_v \text{ of } N_0 \text{ [prep } N_1] \text{ [prep } N_2] \rangle$  si la classe<sub>*i*</sub> contient le schéma  $\langle N_0 \text{ V prep } N_1 \text{ prep } N_2 \rangle$  ainsi que le schéma  $\langle N_0 \text{ V} \rangle$  (lignes 32 à 34).
  6.  $gnp_0$  prend la valeur  $\langle N_v \text{ of } N_0 \text{ prep } N_1 \text{ [prep } N_2] \rangle$  si cette classe ne contient pas le schéma  $\langle N_0 \text{ V} \rangle$  mais contient le schéma  $\langle N_0 \text{ V prep } N_1 \rangle$  (lignes 35 à 36).

**3) Création des super-classes :** En analysant les classes qui ont été précédemment créées et les GNpréd qui leur sont associés, on a remarqué qu'on pouvait regrouper certaines classes et les mettre dans une même super-classe car elles sont associées à des GNpréd similaires. Le but de cette classification est double : (i) nous permettre de regrouper toutes les nominalisations qui possèdent le même comportement syntaxique dans les mêmes super-classes et (ii) créer des super-classes qui nous permettent de connaître la fonction syntaxique de l'argument qui est introduit par la préposition *of* et diminuer ainsi, le problème d'ambiguïté lié à cette préposition (Cf. Section 4.2.4). Ce dernier point sera traité en détail dans la section 4.4.2. En prenant en considération ces deux points, nous avons obtenu six super-classes (qui seront détaillées dans la section 4.4.2) :

- la super-classe 1 contient toutes les nominalisations qui possèdent les GNpréd suivants :
  - $N_v$  of  $N_1$  [prep  $N_2$ ] by  $N_0$  où  $\langle \text{prep } N_2 \rangle$  est optionnel. Ce GNpréd est associé au schéma  $\langle N_0 \text{ V } N_1 \text{ prep } N_2 \rangle$  lorsqu’il est saturé. Dans le cas où  $\langle \text{prep } N_2 \rangle$  est effacé (non saturé), le GNpréd est associé au schéma  $\langle N_0 \text{ V } N_1 \rangle$ .
  - $N_v$  of  $N_0$  : Ce GNpréd est associé au schéma  $\langle N_0 \text{ V} \rangle$ .
  
- la super-classe 2 regroupe toutes les nominalisations dont les GNpréd possèdent les formes :
  - $N_v$  of  $N_1$  [Prep<sub>a</sub>  $N_2$ ] by  $N_0$ . Ce GNpréd est associé au schéma  $\langle N_0 \text{ V } N_1 \text{ prep}_a N_2 \rangle$  lorsqu’il est saturé ainsi qu’au schéma  $\langle N_0 \text{ V } N_1 \rangle$  lorsqu’il est non saturé.
  - $N_v$  of  $N_0$  [Prep<sub>b</sub>  $N_1$ ][Prep<sub>c</sub>  $N_2$ ] : il est associé au schéma  $\langle N_0 \text{ V prep}_b N_1 \text{ prep}_c N_2 \rangle$  lorsqu’il est saturé, au schéma  $\langle N_0 \text{ V prep}_b N_1 \rangle$  lorsque  $\langle \text{prep}_c N_2 \rangle$  est effacé et au schéma  $\langle N_0 \text{ V} \rangle$  lorsque  $\langle \text{prep}_b N_1 \rangle$  et  $\langle \text{prep}_c N_2 \rangle$  sont tous deux effacés. Nous avons choisi d’utiliser les notations prep<sub>a</sub>, prep<sub>b</sub> et prep<sub>c</sub> pour distinguer les prépositions qui peuvent être différentes dans le groupe prépositionnel rattaché au premier schéma nominal (prep<sub>a</sub>) de celles du second schéma (prep<sub>b</sub> et prep<sub>c</sub>).
  
- la super-classe 3 représente deux types de GNpréd :
  - $N_v$  of  $N_1$  by  $N_0$  : associé au schéma  $\langle N_0 \text{ V } N_1 \rangle$ .
  - $N_v$  of  $N_0$  [prep  $N_1$ ] [prep  $N_2$ ] : ce GNpréd est associé à trois schémas verbaux : (i) le premier est associé au schéma  $\langle N_0 \text{ V prep } N_1 \text{ prep } N_2 \rangle$  lorsqu’il est saturé, (ii) le second à un schéma  $\langle N_0 \text{ V prep } N_1 \rangle$  lorsque  $\langle \text{prep } N_2 \rangle$  est effacé, (iii) et à un schéma  $\langle N_0 \text{ V} \rangle$  lorsque  $\langle \text{prep } N_1 \rangle$  et  $\langle \text{prep } N_2 \rangle$  sont tous les deux effacés.
  
- la super-classe 4 représente le GNpréd  $\langle N_v \text{ of } N_1 \text{ [prep } N_2] \text{ by } N_0 \rangle$ . Ce dernier est associé à un schéma verbal de type  $\langle N_0 \text{ V } N_1 \text{ prep } N_2 \rangle$  lorsqu’il saturé et à un schéma  $\langle N_0 \text{ V } N_1 \rangle$  lorsque  $\langle \text{prep } N_2 \rangle$  est effacé.
  
- la super-classe 5 regroupe toutes les nominalisations qui peuvent être représentées par le GNpréd  $\langle N_v \text{ of } N_0 \text{ [prep } N_1] \text{ [prep } N_2] \rangle$  où  $\langle \text{prep } N_1 \rangle$  et  $\langle \text{prep } N_2 \rangle$  sont optionnels. Quand le GNpréd est saturé, c’est-à-dire quand  $\langle \text{prep } N_1 \rangle$  et  $\langle \text{prep } N_2 \rangle$  sont présents, il est associé à un schéma verbal de type  $\langle N_0 \text{ V prep } N_1 \text{ prep } N_2 \rangle$ . Lorsque ces derniers sont effacés, le GNpréd résultant  $\langle N_v \text{ of } N_0 \rangle$  est associé à un schéma  $\langle N_0 \text{ V} \rangle$ . Lorsque

seul  $\langle \text{prep } N_2 \rangle$  est effacé, le GNpréd résultant  $\langle N_v \text{ of } N_0 [\text{prep } N_1] \rangle$  est associé à un schéma de type  $\langle N_0 \text{ V prep } N_1 \rangle$ .

- la super-classe 6 : elle représente le GNpréd  $\langle N_v \text{ of } N_0 [\text{prep } N_1] \rangle$  dans lequel  $\langle \text{prep } N_1 \rangle$  est optionnel. Ce GNpréd est associé à un schéma verbal transitif indirect lorsqu’il est saturé et à un schéma intransitif dans le cas contraire.

Le programme qui nous a permis de créer les super-classes est représenté par l’algorithme 2.

---

**Algorithme 2** : L’algorithme qui permet de créer les super-classes

---

**Entrée** : Les 31 classes de GNpréd  
**Sortie** : Les différentes super-classes

```

1 Début
2   tab-classesN[31] : tableau d’enregistrements; /* défini dans l’algorithme 1 */;
3   mat-superclasses[31][31] : matrice; /* Les lignes de cette matrice représentent
   les numéros des super-classes, les colonnes les numéros des classes et
   l’élément (i,j) l’adresse de la classe */;
4   i, j : entier;
5   j ← 0;
6   Pour i ← 0 à 30 faire
7     Si tab-classesN[i].traite ≠ 1 Alors
8       mat-superclasses[i][i] ← adresse(tab-classesN[i]);
9       tab-classesN[i].traite ← 1;
10      Pour j ← i + 1 à 30 faire
11        Si tab-classesN[j].traite ≠ 1 et
12          equivalent(tab-classesN[i], tab-classesN[j]) Alors
13            mat-superclasses[i][j] ← adresse(tab-classesN[j]);
14            tab-classesN[j].traite ← 1;
15          FinSi
16        FinPour
17      FinSi
18    FinPour
19 Fin

```

---

Ce dernier a pour but de comparer les gnp (gnp0 et gnp1) associés à chaque classe et regrouper celles qui sont associées à des GNpréd qui possèdent un comportement syntaxique similaire dans la même super-classe. Pour cela, on utilise le tableau d’enregistrements *tab-classesN* (ligne 2) qui a été défini dans l’algorithme 1. On déclare aussi *mat-superclasses* (ligne 3), une matrice dans laquelle sont stockés les résultats de la classification. Les lignes de la matrice

---

**Fonction**  $\text{equivalent}(classe_i, classe_j)$ 

---

**Entrée** :  $classe_i, classe_j$

**Sortie** : Vrai si  $classe_i$  et  $classe_j$  sont équivalents, faux sinon

1 **Début**

```
2   Si ( $classe_i.nbreGnp= 1$  et  $classe_j.nbreGnp= 1$ ) Alors
3     Si ( $(classe_i.gnp1 \neq \emptyset$  et  $classe_j.gnp1 \neq \emptyset)$  et
4        $\text{estPresent}("of N_1", classe_i.gnp1, classe_j.gnp1)$ ) Alors
5         retourner vrai;
6     Si ( $(classe_i.gnp0 \neq \emptyset$  et  $classe_j.gnp0 \neq \emptyset)$  et
7        $\text{estPresent}("of N_0", classe_i.gnp0, classe_j.gnp0)$  et  $\neg \text{estPresent}("Prep$ 
8          $N_2", classe_i.gnp0, classe_j.gnp0)$ ) Alors
9           retourner vrai;
10    FinSi
11  Si ( $classe_i.nbreGnp= 2$  et  $classe_j.nbreGnp= 2$ ) Alors
12    Si ( $\neg \text{estPresent}("Prep", classe_i.gnp0, classe_j.gnp0)$  et
13       $\text{estPresent}("Prep", classe_i.gnp1, classe_j.gnp1)$ )
14    ou ( $\neg \text{estPresent}("Prep", classe_i.gnp1, classe_j.gnp1)$ )
15    ou ( $\text{estPresent}("Prep", classe_i.gnp0, classe_j.gnp0)$  et
16       $\text{estPresent}("Prep", classe_i.gnp1, classe_j.gnp1)$ ) Alors
17      retourner vrai
18  FinSi
19  retourner faux ;
20 Fin
```

---

représentent les différentes super-classes, les colonnes les numéros des classes et les éléments (i, j) les adresses des différentes classes. Le principe de cet algorithme est le suivant :

En parcourant le tableau  $tab\text{-}classesN$  (lignes 6 à 17) et pour chaque  $classe_i$ , on effectue les opérations suivantes :

- si l'élément courant ( $classe_i$ ) n'a pas déjà été classé dans une super-classe (ligne 7), on compare ses  $gnp$  ( $gnp_0$  et  $gnp_1$ ) à ceux des autres éléments de  $tab\text{-}classesN$  ( $classe_j$ )

(ligne 10) qui n'ont pas déjà été classés (ligne 11). Si le résultat de la comparaison retourné par la fonction `equivalent` est vrai, on insère l'adresse de la classe<sub>j</sub> à la ligne<sub>i</sub> et à la colonne<sub>j</sub> de la matrice `mat-superclasses[ ][ ]` (ligne 12). On marque ensuite la classe<sub>j</sub> comme étant déjà classée (ligne 13).

La fonction `equivalent` a pour but de renvoyer vrai si les deux classes passées en argument (classe<sub>i</sub> et classe<sub>j</sub>) sont équivalentes ou faux dans le cas contraire. Comme nous l'avons déjà mentionné, nous avons admis dans les algorithmes que `gnp0` correspond au GNpréd sans COD et `gnp1` à ceux avec un COD. Le principe de cette fonction est le suivant :

1. on vérifie si les classes *i* et *j* possèdent les mêmes `gnp`. Si le nombre de `gnp` que possède chaque classe est égal à un (ligne 2) :
  - les tests des lignes 3 à 8 permettent de regrouper les classes qui ne contiennent que les GNpréd associés aux schéma verbaux transitifs avec COD ainsi que les schémas intransitifs sans COD (avec et sans un deuxième complément prépositionnel). Pour cela, on teste si les `gnp1` des deux classes contiennent la chaîne "of N<sub>1</sub>" et si `gnp0` des mêmes classes contiennent les chaînes "of N<sub>0</sub>" et/ou "Prep N<sub>2</sub>". La fonction `equivalent` retourne vrai si une de ces conditions est vraie. Ce test nous a permis d'obtenir les super-classes 4, 5 et 6 (Cf. Figure 4.6).
2. si le nombre de `gnp` de chaque classe est égal à deux :
  - on regroupe les classes dont la chaîne "Prep" n'est pas présente dans les `gnp0` mais présente dans les `gnp1` (ligne 12). Ce test permet d'obtenir la super-classe 1.
  - les classes dont les `gnp1` ne contiennent pas la chaîne "Prep" sont regroupées (ligne 13). Cela nous a permis d'obtenir la super-classe 3.
  - le test de la ligne 14 permet de regrouper les classes dont les `gnp0` et `gnp1` contiennent la chaîne "Prep" (super-classe 2).

#### 4.4.2 Les ambiguïtés liées aux super-classes

Les super-classes ont été créées en utilisant des heuristiques qui nous ont permis de traiter l'ambiguïté relative à la préposition *of* et de la minimiser. Ceci veut dire que la classification a été faite de telle manière que lorsqu'on analyse un GNpréd contenant une nominalisation qui appartient à une super-classe donnée, la probabilité de pouvoir déterminer ce que les marqueurs introduisent comme arguments nominaux soit maximale et que l'ambiguïté soit minimale. Il y a ambiguïté lorsque le même marqueur introduit différents arguments nominaux. Il est intéressant

de noter que tous les cas d'ambiguïtés étudiés dans ce chapitre sont causés principalement par la préposition *of* et son rôle de marqueur, à la fois, du sujet et du COD (Cf. Section 4.2.4). Par exemple, à partir du GNpréd 4.36b on peut observer que la préposition *of* est le marqueur du COD mais le GNpréd 4.37b nous montre que la même préposition est le marqueur du sujet. Le rôle de la préposition *of* devient plus difficile à déterminer lorsque le GNpréd est non saturé. Par exemple, lorsque les postmodificateurs *by Marie* et *for a loan* sont effacés, les GNpréd résultant *the application of a s* et *the application of John* ne permettront pas à l'analyseur de déduire le rôle de la préposition *of* et par conséquent, la fonction syntaxique de l'argument *a poultice/John*.

- (4.36) (a) *Marie applied a poultice*  
 (Marie a appliqué un cataplasme)  
 (b) *the application of a poultice by Marie*  
 (l'application d'un cataplasme par Marie)

- (4.37) (a) *John applied for a loan*  
 (John a fait une demande de prêt)  
 (b) *the application of John for a loan*  
 (la demande de John pour un emprunt)

Dans le tableau 4.2, on détaille toutes les super-classes en associant à chacune d'elle les schémas nominaux qu'elles peuvent admettre et pour chaque schéma nominal, les cas d'ambiguïté et de non-ambiguïté, c'est-à-dire, dans quel cas on peut déterminer les arguments sujet et COD qui peuvent tous les deux être introduits par la préposition *of*.

- la super-classe 1 représente deux GNpréd. Le premier est associé à un schéma  $\langle N_0 V N_1 \rangle$  lorsqu'il est non saturé et à un schéma  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  lorsqu'il est saturé. Dans les deux cas, la préposition *of* introduit le COD et le *by* introduit le sujet. Le second GNpréd est associé à un schéma intransitif ( $N_0 V$ ) dans lequel la préposition *of* introduit le sujet. Il y a ambiguïté lorsque (*prep N<sub>2</sub> ET by N<sub>0</sub>*) sont effacés. Dans ce cas, on ne peut pas savoir si le GNpréd résultant  $\langle N_v \text{ of } N_{0/1} \rangle$  est associé au schéma  $\langle N_0 V \rangle$  ou aux schémas  $\langle N_0 V N_1 \rangle / \langle N_0 V N_1 \text{ prep } N_2 \rangle$ . Par conséquent, il nous est impossible de savoir si la préposition *of* introduit un argument sujet ou objet. Par exemple, la nominalisation *charge* admet trois GNpréd : (i) un GNpréd (exemple 4.38b) associé à un schéma verbal de type  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  (exemple 4.38a), (ii) un GNpréd (exemple 4.38d) associé à un schéma  $\langle N_0 V N_1 \rangle$  (exemple 4.38c), et (iii) un

Superclasses	Schémas nominaux	Cas non ambigus
Superclasse 1 <i>charge, remittance, bid, remission, etc.</i>	$N_V$ of $N_1$ [Prep $N_2$ ] by $N_0$	[by $N_0$ ] OU [Prep $N_2$ ] sont présents
	$N_V$ of $N_0$	
Superclasse 2 <i>alternation, running, spotting, etc.</i>	$N_V$ of $N_1$ [Prep <sub>a</sub> $N_2$ ] by $N_0$	<b>Cas 1 :</b> Prep <sub>a</sub> ≠ Prep <sub>b</sub> ET Prep <sub>a</sub> ≠ Prep <sub>c</sub> [by $N_0$ ] OU [Prep <sub>a/b/c</sub> $N$ ] présent <b>Cas 2 :</b> Prep <sub>a</sub> = Prep <sub>b</sub> [by $N_0$ ] OU [Prep <sub>c</sub> $N_2$ ] présent <b>Cas 3 :</b> Prep <sub>a</sub> = Prep <sub>c</sub> [by $N_0$ ] OU [Prep <sub>b</sub> $N_1$ ] présent
	$N_V$ of $N_0$ [Prep <sub>b</sub> $N_1$ ][Prep <sub>c</sub> $N_2$ ]	
Superclasse 3 <i>variation, presumption, radiation, etc.</i>	$N_V$ of $N_1$ by $N_0$	[by $N_0$ ] OU [Prep $N_1$ ] OU [Prep $N_2$ ] OU ([Prep $N_1$ ] ET [Prep $N_2$ ]) sont présents
	$N_V$ of $N_0$ [Prep $N_1$ ][Prep $N_2$ ]	
Superclasse 4 <i>rendering, respotting, allowance, etc.</i>	$N_V$ of $N_1$ [Prep $N_2$ ] by $N_0$	Dans tous les cas
Superclasse 5 <i>complaint, participation, intercession, etc.</i>	$N_V$ of $N_0$ [Prep $N_1$ ][Prep $N_2$ ]	Dans tous les cas
Superclasse 6 <i>rise, crusade, exclamation, etc.</i>	$N_V$ of $N_0$ [Prep $N_1$ ]	Dans tous les cas

TABLE 4.2 – Les ambiguïtés liées aux super-classes

GNpréd (exemple 4.39b) associé à un schéma verbal intransitif (exemple 4.39a). On a une ambiguïté lorsque  $\langle \text{prep } N_2 \rangle$  (*with a murder*) et  $\langle \text{by } N_0 \rangle$  (*by the judge*) sont effacés. En effet, si l'analyseur analyse les GNpréd *the charge of Paul* et *the charge of the elephant*, qui sont de la forme  $\langle N_v \text{ of } N \rangle$ , il ne pourra pas déterminer les fonctions syntaxiques (sujet ou COD) des arguments *Paul* et *the elephant*.

- (4.38) (a) *the judge charged Paul with a murder*  
 (le juge a inculpé Paul d'un crime)  
 (b) *the charge of Paul with a crime by the judge*  
 (l'inculpation de Paul d'un meurtre par le juge)  
 (c) *the judge charged Paul*  
 (le juge a accusé Paul)  
 (d) *the charge of Paul by the judge*  
 (l'accusation de Paul par le juge)

- (4.39) (a) *the elephant charged*  
 (l'éléphant a chargé)  
 (b) *the charge of the elephant*  
 (la charge de l'éléphant)

- la super-classe 2 admet deux GNpréd : le premier GNpréd ( $N_v \text{ of } N_1 [\text{Prep}_a N_2] \text{ by } N_0$ ) est associé à un schéma verbal de type  $\langle N_0 V N_1 \text{ prep}_a N_2 \rangle$  lorsqu'il est saturé et au schéma  $\langle N_0 V N_1 \rangle$  lorsqu'il est non saturé. Dans les deux cas, la préposition *of* introduit le COD et le *by* introduit le sujet (exemples 4.40a et 4.40b). Le deuxième GNpréd ( $N_v \text{ of } N_0 [\text{Prep}_b N_1] [\text{Prep}_c N_2]$ ) est associé aux schémas : (i)  $\langle N_0 V \text{ prep}_b N_1 \text{ prep}_c N_2 \rangle$  lorsqu'il est saturé (exemples 4.41a et 4.41b), (ii) au schéma  $\langle N_0 V \text{ prep}_b N_1 \rangle$  lorsque  $\langle \text{prep}_c N_2 \rangle$  peut s'effacer ou lorsque le verbe possède un emploi verbal  $\langle N_0 V \text{ prep } N_1 \rangle$  qui admet une préposition différente de  $\langle \text{prep}_b \rangle$  (exemples 4.41c et 4.41d) et (iii) au schéma  $\langle N_0 V \rangle$  lorsque  $\langle \text{prep}_b N_1 \rangle$  et  $\langle \text{prep}_c N_2 \rangle$  sont effacés (exemples 4.41e et 4.41f). On remarque dans les exemples 4.41a et 4.41b qu'il n'est pas possible de passer du GNpréd  $\langle N_V \text{ of } N_0 [\text{prep}_b N_1] [\text{prep}_c N_2] \rangle$  au GNpréd  $\langle N_V \text{ of } N_0 [\text{prep}_b N_1] \rangle$  (\* *the alternation of the color from black*). Ceci est dû au fait que les prépositions *from* et *to* forment un bloc et sont dans la très grande majorité des cas indissociables. Par contre, il est possible d'avoir le GNpréd  $\langle N_V \text{ of } N_0 \rangle$  (*the alternation of the color*) si les deux arguments *from black* et *to white* sont effacés ensemble. Dans

tous ces GNpréd, la préposition *of* introduit le sujet. Plusieurs cas d'ambiguïtés peuvent se présenter : lorsque  $\langle \text{by } N_0 \rangle$  ou encore  $\langle \text{prep}_c N_2 \rangle$  sont effacés, dans la situation où  $\text{prep}_a = \text{prep}_b$ . Dans ce cas, le GNpréd résultant  $\langle N_v \text{ of } N_{0/1} \text{ prep}_{b/a} N_{1/2} \rangle$  ne nous permet pas de savoir s'il est associé à un schéma verbal de type  $\langle N_0 \text{ V } N_1 \text{ Prep } N_2 \rangle$  ou s'il est associé à un schéma verbal de type  $\langle N_0 \text{ V } \text{Prep } N_1 \rangle$ . En d'autres termes, on ne peut pas déterminer si la préposition *of* introduit le COD ou le sujet. Un autre cas d'ambiguïté a lieu lorsque  $(\text{by } N_0 \text{ ET } \text{prep}_a N_2)$  pour le premier schéma ou  $(\text{prep}_b N_1 \text{ ET } \text{prep}_c N_2)$  pour le second schéma sont effacés. Dans ce cas aussi, on ne peut pas savoir à quel schéma verbal le GNpréd résultant  $\langle N_v \text{ of } N_{0/1} \rangle$  est associé. Pour les mêmes raisons, on ne peut pas savoir si la préposition *of* introduit le COD ou le sujet.

- (4.40) (a) *John alternated alcohol with another beverage*  
 (John a alterné l'alcool avec une autre boisson)  
 (b) *the alternation of the alcohol with another beverage by John*  
 (l'alternance de l'alcool avec une autre boisson par John)  
 (c) *John alternates comedy acts*  
 (John alterne les numéros de comédie)  
 (d) *the alternation of the comedy acts by John*  
 (l'alternance des numéros de comédie par John)

- (4.41) (a) *the color alternates from black to white*  
 (la couleur alterne du noir au blanc)  
 (b) *the alternation of the color from black to white*  
 (l'alternance de la couleur du noir au blanc)  
 (c) *wet days alternated with fine days*  
 (les jours pluvieux alternaient avec les beaux jours)  
 (d) *the alternation of wet days with fine days*  
 (l'alternance des jours pluvieux avec les beaux jours)  
 (e) *generations alternate*  
 (les générations alternent)  
 (f) *the alternation of generations*  
 (l'alternance des générations)

– la super-classe 3 admet deux GNpréd : le premier GNpréd est associé au schéma  $\langle N_0 \text{ V } N_1 \rangle$  dans lequel le COD et le sujet sont introduits respectivement

par les prépositions *of* et *by* (exemples 4.42a et 4.42b). Le deuxième GNpréd ( $N_v$  of  $N_0$  [Prep  $N_1$ ] [Prep  $N_2$ ]) est associé aux schémas : (i)  $\langle N_0 V \text{ prep } N_1 \text{ prep } N_2 \rangle$  lorsqu'il est saturé (exemples 4.43a et 4.43b), (ii)  $\langle N_0 V \text{ prep } N_1 \rangle$  lorsque  $\langle \text{prep } N_2 \rangle$  peut s'effacer ou lorsque le verbe possède un emploi verbal  $\langle N_0 V \text{ Prep } N_1 \rangle$  qui admet une préposition différente de celle qui introduit  $N_2$  dans le schéma  $\langle N_0 V \text{ prep } N_1 \text{ prep } N_2 \rangle$  (exemples 4.43c et 4.43d) et (iii)  $\langle N_0 V \rangle$  lorsque *prep*  $N_1$  et *prep*  $N_2$  sont effacés (exemples 4.43a et 4.43b). On remarque dans les exemples 4.43a et 4.43b qu'il n'est pas possible de passer du GNpréd  $\langle N_V \text{ of } N_0 \text{ [prep } N_1] \text{ [prep } N_2] \rangle$  au GNpréd  $\langle N_V \text{ of } N_0 \text{ prep } N_1 \rangle$  (*\*the variation of the position from 0°*) car, dans ce cas, les prépositions *from* et *to* forment un bloc et sont dans la très grande majorité des cas indissociables. Par contre, il est possible d'avoir le GNpréd  $\langle N_V \text{ of } N_0 \rangle$  (*the variation of the position*) si *from 0°* et *to 30°* sont effacés ensemble. Dans tous ces cas, la préposition *of* introduit le sujet. Le seul cas d'ambiguïté se produit lorsque (*by*  $N_0$  ET *prep*  $N_1$  ET *prep*  $N_2$ ) sont effacés. Dans ce cas, on ne pourra pas savoir si le GNpréd résultant  $\langle N_v \text{ of } N_{0/1} \rangle$  est associé à un schéma verbal transitif indirect/intransitif ou à un schéma verbal transitif direct. Par conséquent, il est difficile de déterminer si le *of* introduit le sujet ou le COD. Ainsi, les deux GNpréd *the variation of the decoration* et *the variation of life expectancy* ne permettent pas à l'analyseur de déterminer si les arguments *the decoration* et *the behavior* sont sujet ou COD.

(4.42) (a) *Marie varied the decoration*

(Marie a varié la décoration)

(b) *the variation of decoration by Marie*

(la variation de la décoration par Marie)

(4.43) (a) *The position varied [from 0° to 30°]*

(la position a varié de 0° à 30°)

(b) *the variation of the position [from 0° to 30°]*

(la variation de la position [de 0° à 30°])

(c) *life expectancy varies between men and women*

(l'espérance de vie varie entre les hommes et les femmes)

(d) *the variation of life expectancy between men and women*

(la variation de l'espérance de vie varie entre les hommes et les femmes)

– la super-classe 4 n'admet qu'un seul GNpréd. Ce dernier est associé à un schéma verbal

de type  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  lorsqu'il est saturé (avec  $\text{prep } N_2$  présent) et à un schéma de type  $\langle N_0 V N_1 \rangle$  lorsque  $\text{prep } N_2$  est effacé (exemple 4.44a et 4.44b). Dans les deux cas, la préposition *of* introduit le COD et par conséquent, les GNpréd qui appartiennent à cette super-classe ne peuvent pas être ambigus. Par exemple, même en supprimant les groupes prépositionnels *into English* et *by Nicolas* du GNpréd 4.44b, l'analyseur pourra toujours déterminer que la préposition *of* est l'introducteur du COD.

- (4.44) (a) *Nicolas rendered the document [into English]*  
 (Nicolas a traduit le document [en anglais])  
 (b) *the rendering of the document [into English] by Nicolas*  
 (la traduction du document [en anglais] par Nicolas)

- la super-classe 5 n'admet que le GNpréd  $\langle N_V \text{ of } N_0 [\text{Prep } N_1] [\text{Prep } N_2] \rangle$ . Ce GNpréd est associé à trois types de schémas verbaux : (i) à un schéma verbal de type  $\langle N_0 V \text{ prep } N_1 \text{ prep } N_2 \rangle$  lorsqu'il est saturé ( $\text{prep } N_1$  et  $\text{prep } N_2$  sont présents), (ii) à un schéma verbal de type  $\langle N_0 V \text{ prep } N_1 \rangle$  lorsque  $\text{prep } N_2$  est effacé, et (iii) au schéma  $\langle N_0 V \rangle$  lorsque  $\text{prep } N_1$  et  $\text{prep } N_2$  sont effacés (exemples 4.45a et 4.45b). Dans tous les cas, la préposition *of* introduit un sujet et par conséquent, les GNpréd qui appartiennent à cette super-classe ne peuvent pas être ambigus.

- (4.45) (a) *a person complained [to the police chief] [about the conduct of . . .]*  
 (une personne s'est plainte [au chef de la police] [au sujet de la conduite de . . .])  
 (b) *the complaint of a person [to the police chief] [about the conduct of . . .]*  
 (la plainte d'une personne [au chef de la police] [au sujet de la conduite de . . .])

- la super-classe 6 n'admet qu'un seul GNpréd. Sa forme saturée (exemple 4.46b) est associée à un schéma de type  $\langle N_0 V \text{ prep } N_1 \rangle$  (exemple 4.46a) et sa forme non saturée (exemple 4.47b) est associée au schéma  $\langle N_0 V \rangle$  (exemple 4.47a). Dans tous les cas de figures, la préposition *of* introduit le sujet et par conséquent, les GNpréd qui appartiennent à cette super-classe ne peuvent pas être ambigus.

- (4.46) (a) *the temperature rises above zero*  
 (la température est montée au-dessus de zéro)  
 (b) *the rise of the temperature above zero*  
 (la montée de la température au-dessus de zéro)

- (4.47) (a) *heat rises*  
(la chaleur monte)  
(b) *the rise of heat*  
(la montée de chaleur)

### 4.4.3 Constitution du lexique de nominalisations

L'étape suivante consiste à regrouper les nominalisations. Pour cela, on a associé à chaque nominalisation les différentes prépositions qu'elle peut admettre dans les différents GNpréd qu'ils lui correspondent. À l'aide des deux modules BD et PPLC (Cf. Section 4.3.2) nous avons : (i) récupéré toutes les nominalisations qui appartiennent à une super-classe donnée, (ii) regroupé les nominalisations suivant les prépositions qu'elles admettent, (iii) associé chaque nominalisation aux différentes prépositions qu'elle admet, et (iv) associé chaque nominalisation au schéma syntaxique du/des GNpréd qu'elle admet. En d'autres termes, on associe à chaque nominalisation une information qui renseigne sur les différentes prépositions admises par la nominalisation ainsi que sur les schémas syntaxiques des GNpréd admis. Il est possible, à cette étape, de générer des fichiers de nominalisations qui satisfassent des contraintes linguistiques particulières. Nous montrons ici comment nous avons généré des fichiers répondant aux contraintes d'une grammaire ajoutée à la grammaire initiale du Link Parser et permettant d'analyser les GNpréd en mettant en évidence leurs liens argumentaux. Par exemple, la nominalisation *abidance* appartient à la super-classe 3. Elle est donc associée à trois types de GNpréd associés aux schémas verbaux suivants :  $\langle N_0 V N_1 \rangle$ ,  $\langle N_0 V \rangle$  et  $\langle N_0 V \text{ prep } N_1 \rangle$  avec les prépositions *at*, *in*, *with* et *by*. On mettra cette nominalisation dans deux fichiers distincts : (i) un fichier qui permet d'analyser des GNpréd reliés à un schéma transitif direct et (ii) un autre, capable d'analyser des GNpréd reliés aux schémas verbaux  $\langle N_0 V \text{ prep } N_1 \rangle$  pour chacune des 3 prépositions. De plus, des fichiers spéciaux sont générés pour signaler les cas où le rattachement argumental est ambigu dans les cas d'effacement.

Il est possible ainsi d'obtenir un lexique de nominalisations correspondant aux types syntaxiques des schémas nominaux et aux différentes prépositions utilisées, si ces schémas en admettent.

### 4.4.4 Traitement des entrées lexicales complexes

Certaines nominalisations possèdent des schémas nominaux dont les structures syntaxiques et fonctionnelles ne peuvent pas être déduites automatiquement à partir des schémas verbaux qui leur sont associés. Les nominalisations que nous avons examinées jusqu'à présent disposent de

schémas nominaux dont la déduction peut se faire en se basant sur les schémas verbaux qui leur sont associés. Par contre, les schémas nominaux de certaines entrées lexicales se caractérisent par des propriétés syntaxiques particulières qui nécessitent un traitement semi-automatique. Par exemple, PredicateDB nous a permis de trouver le cas du verbe *absorb* et sa nominalisation *absorption*. Dans le SL, *absorb* n'admet que schéma verbal transitif direct. Comme on l'a vu (Tableau 4.1), le schéma nominal associé à ce schéma admet un COD et un sujet introduits respectivement par les prépositions *of* et *by*. Or, la description de la nominalisation *absorption* dans le SL montre que ses compléments de nom sont introduits par les prépositions *of*, *by*, *in* et *into*. En analysant les exemples 4.48a, 4.48b et 4.48c, on peut déduire que la préposition *of* introduit toujours le COD, mais que dans certains cas, la préposition *by* peut être substituée par les prépositions *in* et *into* (Schéma 4.49).

- (4.48) (a) *the absorption of the proteins by the [cells]<sub>sujet</sub>*  
 (b) *the absorption of the proteins in the [cells]<sub>sujet</sub>*  
 (c) *the absorption of the proteins into the [cells]<sub>sujet</sub>*  
 (l'absorption des protéines par/dans les cellules)

$$N_0 \text{ absorb } N_1 \Leftrightarrow \text{the absorption of } [N_1]_{\text{COD}} \text{ by/in/into } [N_0]_{\text{Sujet}} \quad (4.49)$$

PredicateDB nous a permis d'analyser une autre classe de prédicats nominaux dits prédicats nominaux symétriques. Ils sont appelés ainsi car ils sont dérivés de verbes symétriques. Les nominalisations symétriques sont intéressantes car elles sont très productives dans le sous-langage de la génomique. Un prédicat (verbal/nominal) est dit symétrique lorsque ses arguments peuvent être permutés sans que cela n'affecte le sens du groupe verbal ou du GNpréd correspondant. Par exemple, le couple *abide/abidance* (séjourner/séjour) sont symétriques comme le montrent les exemples 4.50 et 4.51. En effet, on remarque qu'il est possible de permuter les arguments *Régis* et *Jérôme* sans que cela n'altère le sens de la phrase ou du GNpréd.

- (4.50) (a) *Jérôme abode with Régis in the same hotel*  
 (Jérôme a séjourné avec Régis dans le même hôtel)  
 (b) *the abidance of Jérôme with Régis in the same hotel*  
 (le séjour de Jérôme avec Régis dans le même hôtel)

- (4.51) (a) *Régis abode with Jérôme in the same hotel*  
 (Régis a séjourné avec Jérôme dans le même hotel)  
 (b) *the abidance of Régis with Jérôme in the same hotel*  
 (le séjour de Régis avec Jérôme dans le même hotel)

En se basant sur les différentes formes des schémas syntaxiques admis par les nominalisations symétriques, on peut déduire différentes classes ou modèles :

1. Les verbes/nominalisations de la première classe possèdent des schémas syntaxiques qui se composent de deux arguments ou d'un ensemble d'arguments qui appartiennent à la même famille et qui jouent le rôle du sujet. Les patrons syntaxiques de ces verbes/nominalisations sont représentés par les schémas 4.52, 4.53 et 4.54. Dans les deux premiers schémas, les arguments reçoivent une notation particulière :  $N_a$  et  $N_b$  pour signifier que chacun de ces arguments peut être permuté sans que cela n'altère le sens de la phrase ou du GNpréd. Par exemple, on observe dans l'exemple associé au schéma 4.52 que chacun des arguments (*kinesin* ou *calmodulin*) peut être sujet ou objet ( $N_a \text{ interacts with } N_b \Leftrightarrow N_b \text{ interacts with } N_a$ ). Le deuxième schéma 4.53, qui est strictement équivalent au premier, montre que le sujet peut être composé des deux arguments reliés par conjonction de coordination ( $N_a \text{ and } N_b : \text{Alpha 4 and midi}$ ). Dans ce schéma, la permutation se fait de la façon suivante :  $N_a \text{ and } N_b \text{ interact} \Leftrightarrow N_b \text{ and } N_a \text{ interact}$ . Dans le schéma 4.54, le sujet fait référence dans la plupart des cas à un nom pluriel (noté  $N_{plur}$ ) qui représente un ensemble d'éléments appartenant à la même classe générique (exemples 4.54a et 4.54b). Par exemple, l'argument *EGF receptor family members* de l'exemple 4.54 fait référence à différents membres faisant tous partie de la même classe générique qui est la famille des récepteurs EGF.

- (4.52) -  $N_a \text{ V with } N_b \Leftrightarrow N_v \text{ of } N_a \text{ with } N_b$   
 (a) *kinesin interacts with calmodulin*  $\leftrightarrow$  *interaction of kinesin with calmodulin*  
 (la kinésine interagit avec la calmoduline  $\leftrightarrow$  l'interaction de la kinésine avec la calmoduline)

- (4.53)  $\equiv N_a \text{ and } N_b \text{ V} \leftrightarrow N_v \text{ of/between } N_a \text{ and } N_b$   
 (a) *Alpha 4 and MIDI interact*  $\leftrightarrow$  *the interaction of Alpha 4 and MIDI*  
 (Alpha 4 et MIDI interagissent  $\leftrightarrow$  l'interaction de Alpha 4 et MIDI)

(4.54) -  $N_{plur} V \leftrightarrow N_v$  of/between  $N_{plur}$

(a) *the components interact*  $\leftrightarrow$  *interaction of/between the components*

(les composants interagissent  $\leftrightarrow$  l'interaction des/entre les composants)

(b) *the EGF receptor family members interact*  $\leftrightarrow$  *interaction between EGF receptor family members*

(les membres de la familles des récepteurs EGF interagissent  $\leftrightarrow$  l'interaction entre les membres de la familles des récepteurs EGF)

Pour identifier ce type de nominalisations, nous avons utilisé des requêtes SQL qui nous ont permis d'extraire tous les verbes qui possèdent un schéma verbal de type  $\langle N_0 V \text{ Prep } N_1 \rangle$  admettant uniquement un complément prépositionnel introduit par la préposition *with*. Ces verbes peuvent admettre également un schéma intransitif car ce dernier ne fausse pas les résultats. Ensuite, nous avons croisé ces informations avec les nominalisations qui admettent les prépositions *of*, *with* et *between*. Nous avons de cette façon, mis en évidence un ensemble de nominalisations qui partagent ces propriétés.

2. La seconde classe (modèle) de nominalisations se différencie par rapport à la classe précédente par la présence d'un argument sujet  $N_0$  qui n'est pas concerné par la permutation des arguments. L'argument  $N_0$  est généralement effacé dans le GNpréd bien qu'il puisse être présent dans le schéma verbal qui lui correspond comme le montrent les exemples 4.55 à 4.57 pages 115–116. Le premier schéma montre que les arguments  $N_a$  et  $N_b$  (respectivement *LD* et *CIA*) peuvent permuter (*association of  $N_a$  with  $N_b$  by  $N_0$*   $\leftrightarrow$  *association of  $N_b$  with  $N_a$  by  $N_0$* ). Le deuxième schéma montre que  $N_a$  et  $N_b$  sont reliés par une conjonction de coordination (*focusing nozzle and a rigid macrocarrier*) et que la permutation se fait à l'intérieur du COD (*The association of/between  $N_a$  and  $N_b$  by  $N_0$*   $\leftrightarrow$  *The association of/between  $N_b$  and  $N_a$  by  $N_0$* ). Dans le dernier schéma, la notation  $N_{plur}$  (*constructs*) indique un nom pluriel qui représente un ensemble d'éléments appartenant à la même classe générique comme nous avons pu le voir pour la classe précédente.

(4.55) -  $N_0 V N_a \text{ with } N_b \leftrightarrow N_v$  of  $N_a$  with  $N_b$  by  $N_0$

– *the association of LD with CIA was . . .*

(l'association du LD avec la CIA était . . .)

(4.56)  $\equiv N_0 V N_a \text{ and } N_b \leftrightarrow N_v$  of/between  $N_a$  and  $N_b$  by  $N_0$

- *the association of/between halothane and oxygenated solvents by H NMR spectroscopy*  
(l'association de/entre l'halothane et les solvants oxygénés par la spectroscopie H NMR)

- (4.57)  $\equiv N_0 V N_{plur} \leftrightarrow N_v \text{ of/betwenn } N_{plur} \text{ by } N_0$   
 – *association between constructs in this model is . . .*  
 (l'association entre les différents concepts dans ce modèle est . . .)

Nous avons utilisé les mêmes heuristiques pour identifier les nominalisations appartenant à ce modèle. En utilisant une requête SQL, nous avons extrait tous les verbes qui possèdent un schéma ditransitif direct admettant la préposition *with* et croiser cet ensemble avec les nominalisations correspondantes dont les GNpréd admettent les prépositions *of*, *by* et *with*.

#### 4.4.5 Les propriétés non traitées

Le lexique PredicateDB ainsi créé nous permet de traiter la quasi-totalité des nominalisations verbales que contient le Specialist Lexicon (3 974 nominalisations), ce qui fait que seulement 8 nominalisations (dont les verbes associés n'admettent que des schémas phrastiques) n'ont pas été traitées. Dans notre travail, nous nous sommes concentré uniquement sur les nominalisations associées à des verbes dont les schémas admettent des arguments nominaux. Néanmoins, beaucoup de verbes admettent, en plus de ce type d'arguments, des schémas dont un ou plusieurs arguments sont phrastiques. Par exemple le verbe *promise* peut admettre le schéma verbal  $\langle N_0 V N_1 \text{ prep } N_2 \rangle$  où  $N_1$  et  $N_2$  sont des arguments nominaux (exemple 4.58a). Le GNpréd associé est représenté par l'exemple 4.58b. Ce même verbe peut se construire dans un schéma qui admet un complément phrastique représenté par une infinitive  $\langle N_0 V \text{ to } V_{inf} W \rangle$ <sup>6</sup> (exemple 4.59a). Le complément phrastique (*to bring short term credit . . .*) se retrouve dans le GNpréd associé (exemple 4.59b).

- (4.58) (a) *John promised a job to Micheline*  
 (John a promis un travail pour Micheline)  
 (b) *the promise of a job to Micheline by John*  
 (la promesse d'un travail pour Micheline par John)

---

6. Nous reprenons la terminologie de Gross et désignons par *W* une suite quelconque, éventuellement vide, de compléments. *V<sub>inf</sub>* désigne les verbes à l'infinitif et *V<sub>ing</sub>* ceux au participe présent

- (4.59) (a) *Spain's government promised to bring short term credit under control*  
 (le gouvernement espagnol a promis de ramener le crédit à court terme sous contrôle)  
 (b) *the promise of Spain's government to bring short term credit under control*<sup>7</sup>  
 (la promesse du gouvernement espagnol de ramener le crédit à court terme sous contrôle)

Les schémas qui admettent des compléments phrastiques nécessitent une étude linguistique fine des différentes possibilités d'insertion de ce type de compléments dans la structure du GNpréd. Nous présentons ci-dessous un sous-ensemble de GNpréd acceptant des compléments phrastiques que nous subdivisons en compléments correspondant à des propositions incomplètes d'une part et complètes d'autre part.

**Les propositions incomplètes :** ce type de compléments se construit principalement autour d'une infinitive ou d'un participe présent. Dans la majorité des cas, ces propositions admettent une structure à contrôle, c'est-à-dire que le sujet associé à la proposition (qui n'est pas apparent) est identifié à partir de la proposition principale. Ainsi dans l'exemple 4.59 ci-dessous, le sujet de la proposition *to bring short term credit under control* est celui de la proposition principale (*Spain's government*). L'exemple 4.60 qui est associé au schéma < N<sub>0</sub> V N<sub>1</sub> (to) Ving W > admet un complément construit autour du participe présent (*having*). Dans cet exemple, on remarque un contrôle sur l'objet *each animal* car ce dernier est le sujet de la proposition incomplète *to having its legs . . . .*

- (4.60) (a) *a person habituated each animal to having its legs and body touched*  
 (une personne a habitué chaque animal à avoir ses pattes et son corps touchés)  
 (b) *the habituation of each animal to having its legs and body touched by a person*<sup>8</sup>  
 (l'habituation de chaque animal à avoir ses jambes et son corps touchés par une personne)

**Les propositions complètes :** elles sont définies dans [Browne et al., 2000] comme des propositions qui admettent un verbe conjugué en accord avec le sujet. Il existe différents types de propositions complètes :

– Les complétives : elles sont généralement introduites par la conjonction *that*. Cette dernière

7. [www.creditwritedowns.com/2010/08/spains-national-addiction-to-dinero-b.html](http://www.creditwritedowns.com/2010/08/spains-national-addiction-to-dinero-b.html)

8. <http://www.grandin.com/references/abstract-10.html>

est optionnelle, c'est-à-dire qu'il peut être présent ou non (exemple 4.61a). Dans les deux cas, le *that* est conservé dans les GNpréd associés (exemples 4.61b). Ce type de schémas est représenté dans le Specialist Lexicon par le trait *fincomp* suivi d'une option selon que *that* est présent ou non. C'est ainsi que le verbe *decide* de l'exemple 4.61a est associé, dans le Specialist Lexicon, au schéma verbal *tran=fincomp(o)* où le symbole *o* signifie que *that* est optionnel.

- (4.61) (a) *Tariff Board decided (that) "Clorox" is properly classifiable under tariff item 219a*  
(La commission tarifaire a décidé que le "Clorox" est correctement classifiable sous l'article tarifaire 219a)
- (b) *the decision of Tariff Board that "Clorox" is properly classifiable under tariff item 219a*<sup>9</sup>  
(la décision de la commission tarifaire que le "Clorox" est correctement classifiable sous l'article tarifaire 219a)

– Les formes en WH : ce sont des propositions complètes introduites par un mot qui commence par *WH* (que nous appellerons *WH-mot*) : *whether, where, who*, etc. (exemple 4.62a). Ce type de schémas est noté dans le Specialist Lexicon par le trait *whfincomp*. Les GNpréd associés à ces schémas se caractérisent par le fait qu'ils conservent le *WH-mot* déjà présent dans le schéma verbal (exemple 4.62b).

- (4.62) (a) *I decided how I would go about it* [Browne et al., 2000]  
(J'ai décidé de comment je pourrais m'y prendre)
- (b) *my decision how I would go about it*  
(ma décision de comment je pourrais m'y prendre)

– Les formes infinitives en WH : ces propositions sont notées également dans le Specialist Lexicon par le trait *whinfcomp*. Ce type de compléments requiert un contrôle arbitraire sur le sujet (noté *arbc*), ce qui signifie que le sujet de la subordonnée n'est pas obligatoirement celui de la principale (exemple 4.63). Dans les GNpréd, ce type de compléments est conservé (exemple 4.63b).

- (4.63) (a) *they decided whether to use informal action*  
(ils ont décidé de s'il fallait recourir à une mesure officieuse)

---

9. <http://esc.lexum.umontreal.ca/en/1961/1961scr0-170/1961scr0-170.html>

- (b) *their decision whether to use informal action*<sup>10</sup>  
(leur décision de s'il fallait recourir à une mesure officielle)

Ces quelques exemples choisis parmi les nominalisations qui admettent ce type de compléments ont pour but de montrer l'intérêt de leur insertion dans la structure du GNpréd. Notons que ces propriétés ont été introduites dans la base de données, il reste cependant à décider au cas par cas la possibilité de leur insertion dans la structure du GNpréd.

## 4.5 Le lexique PredicateDB pour l'acquisition de structures argumentales

Ma contribution dans ce travail a été de créer le lexique PredicateDB. Nous présentons dans la suite du document les différentes sous-classes de ce lexique qui a permis l'écriture de la grammaire des GNpréd ainsi que différentes analyses produites. Cette dernière a été intégrée dans la grammaire de l'anglais du Link Parser dans le but de rendre possible l'analyse syntaxique des GNpréd, particulièrement dans le domaine biologique [Royauté et al., 2006, 2007]. Ce travail a permis ensuite de développer PredXtract [Godbert and Royauté, 2009] qui permet d'extraire les arguments des structures prédicatives verbales et nominales d'une phrase. Cette plate-forme : (1) effectue un ensemble d'analyses et sélectionne la meilleure d'entre elles en se basant sur des heuristiques qui privilégient principalement la saturation du GNpréd en reliant la totalité des arguments au nom de tête, et (2) produit ensuite la représentation de cette phrase en un ensemble de structures complexes prédicats-arguments. Pour cela, PredXtract intègre : (i) le Link Parser et sa grammaire (Link Grammar), (ii) une grammaire de liens définie pour l'analyse des structures prédicatives nominales qui permet d'identifier les arguments nominaux qui sont introduits par différentes prépositions, (iii) un module d'alignement entre les structures verbales et nominales, (iv) un module d'extraction des arguments prédicatifs et de sélection des meilleures analyses.

### 4.5.1 Le Link Parser et les grammaires de liens

Le Link Parser est un analyseur syntaxique qui a été développé par Sleator and Temperley [1991]. Le résultat de l'analyse d'une phrase correspond à un graphe dont les arcs étiquetés relient des paires de mots qui sont dans une relation de dépendance. Les auteurs ont mis à la disposition de la communauté une grammaire et un dictionnaire comprenant 60 000 formes de mots couvrant une large variété de constructions syntaxiques de l'anglais. Enfin, le Link Parser

---

10. <http://www.justice.gc.ca/eng/pi/yj-jj/res-rech/discre/situ-conj/situ-conj.html>

mots	formules
a the	D+
enzymes.n compounds.n	{A-} & {D-} & (O- or S+)
inhibite.v	S- & O+
big.a carcinogenic.a	A+

TABLE 4.3 – Comment certains mots sont définis dans le dictionnaire de la grammaire

a la réputation d’être un analyseur robuste. Il est capable d’ignorer un ou plusieurs mots en cas de problèmes d’analyse et d’affecter une structure syntaxique au reste de la phrase. Il peut également traiter des mots inconnus et prédire leurs catégories syntaxiques à partir du contexte syntaxique et de leurs morphologies. Il sait gérer les expressions numériques ainsi que les symboles de ponctuation. Le formalisme grammatical du Link Parser repose sur le formalisme des grammaires de liens qui sont une variante des grammaires de dépendances [Mel’čuk, 1988]. Cette variante grammaticale se caractérise par le fait que les arcs qui lient les paires de mots constituent un graphe planaire dans lequel les liens ne se croisent pas. Une grammaire de liens consiste en un ensemble de mots (les symboles terminaux de la grammaire), où chacun des mots possède des contraintes de connexion. Une séquence de mots constitue une phrase appartenant à la grammaire si les différents liens qui unissent les mots satisfont les conditions suivantes :

1. planarité : les liens ne doivent pas se croiser.
2. connectivité : tous les mots de la phrase doivent être connectés.
3. satisfaction : les liens doivent satisfaire les contraintes de connexion de chaque mot appartenant à la séquence.

Les contraintes de connexion sont définies dans la grammaire qui rassemble les mots ayant les mêmes contraintes. La table 4.3 représente une grammaire simple de quelques mots ainsi que les contraintes de connexion qui leur sont associées. Deux mots peuvent se connecter s’ils ont le même type de connecteur. Par exemple, les déterminants *a* et *the* possèdent un connecteur (D) orienté vers la droite (D+). Ils ne peuvent se connecter qu’avec les mots qui possèdent un connecteur D orienté vers la gauche (D-).

La formule associée aux noms *enzymes* et *compounds* utilise les opérateurs & et or. Elle se compose de trois sous-formules. La première sous-formule représente un connecteur de type A- qui est optionnel (l’optionnalité est exprimée par les accolades). La deuxième sous-formule représente également un connecteur optionnel de type D- et la troisième sous-formule signifie

que ces noms admettent soit un connecteur de type O- soit un connecteur de type S+. La formule générale signifie donc que chacun de ces noms admet un connecteur de type A- (qui peut ne pas être présent), d'un connecteur optionnel de type D- et d'un connecteur de type O- ou S+.

Le verbe *inhibite* admet deux connecteurs : un connecteur S orienté vers la gauche (S-) et un connecteur O orienté vers la droite (O+), reliant respectivement un élément de tête sujet ou objet.

Les adjectifs *big* et *carcinogenic* admettent un connecteur A orienté vers la droite. Ils ne peuvent se connecter qu'avec les mots qui admettent un connecteur A orienté vers la gauche (A-).

La figure 4.7 montre que la phrase *carcinogenic compounds inhibite enzymes*<sup>11</sup> appartient bien à la grammaire car toutes les conditions de connexion des différents mots sont satisfaites.

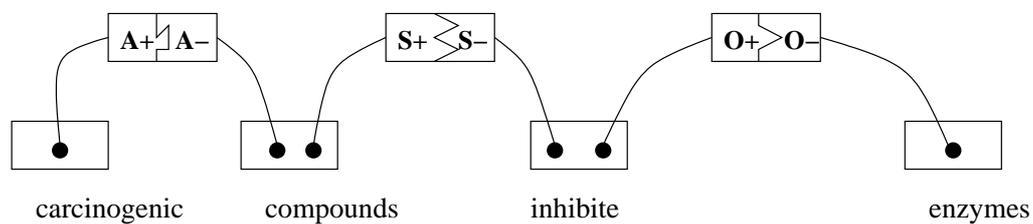


FIGURE 4.7 – Une phrase dans laquelle les contraintes de connexion sont satisfaites

## 4.5.2 Une grammaire pour les prédicats nominaux

L'analyse d'un GNpréd en utilisant la grammaire native du Link Parser ne permet pas de faire apparaître les actants des prédicats nominaux. Pour pouvoir capturer ce type d'arguments et distinguer les compléments essentiels des compléments circonstanciels qui peuvent apparaître dans un GNpréd, il est nécessaire d'enrichir cette grammaire en définissant un ensemble de liens spécifiques qui sont nommés liens argumentaux. Ces liens dépendent des catégories syntaxiques des prédicats nominaux ainsi que des prépositions qui font partie des entrées lexicales de chaque prédicat nominal. Ces informations sont disponibles dans le lexique de PredicateDB.

### Adaptation du lexique PredicateDB pour le Link Parser

Le lexique que nous avons créé est un lexique générique qui peut être utilisé par différents outils linguistiques mais il a été adapté pour être utilisé par le Link Parser. Nous attachons à chaque entrée (nominalisation) appartenant au lexique une extension qui marque le schéma syntaxique qui lui est associé ainsi que les prépositions qui font partie de ses entrées lexicales.

11. Les composés cancérogènes inhibent les enzymes

Lorsqu'une nominalisation possède plusieurs schémas syntaxiques ou prépositions, elle apparaît dans le lexique avec chaque fois une extension différente. Nous donnons ci-dessous un sous-ensemble des principales extensions utilisées :

- nt0 : correspond aux nominalisations associées aux verbes transitifs direct ( $N_0$  V  $N_1$ ). *inhibition.nt0*, *production.nt0* et *accumulation.nt0* en sont un sous-ensemble.
- niX : cette extension est associée aux schémas syntaxiques transitifs indirects  $\langle N_0$  V prep  $N_1 \rangle$ . X représente un chiffre correspondant à un des prépositions de ce schéma et faisant partie des entrées lexicales du prédicat. Par exemple, les nominalisations *response*, *action* et *heterodimerization* possèdent respectivement les extensions ni2, ni3 et ni4 car elles admettent les prépositions *to*, *on* et *with*.
- ndtX : nous associons cette extension aux nominalisations dérivées des verbes qui se construisent avec les schémas qui admettent un COD suivi d'un complément prépositionnel ( $N_0$  V  $N_1$  prep  $N_2$ ). Le symbole X a la même signification que pour l'exemple précédent. Ainsi, les nominalisations *release*, *predisposition* et *protection* possèdent respectivement les extensions ndt3, ndt4 et ndt5 car les prépositions *from*, *to* et *against* font partie de leurs entrées lexicales.
- ns1 : correspond aux nominalisations liées aux verbes symétriques à deux arguments ( $N_a$  V with  $N_b$ ). Les nominalisations telle que *coimmunoprecipitation*, marquées par cette extension sont associées au schéma nominal  $\langle N_v$  of  $N_a$  with  $N_b \rangle$ . Cette nominalisation est également associée au schéma nominal équivalent  $\langle N_v$  of/between  $N_a$  and  $N_b \rangle$  (dérivé du schéma  $\langle N_a$  and  $N_b$  V  $\rangle$ ). Dans ce cas, on lui associe l'extension ns2.
- nsd2 : associée aux nominalisations qui sont liées à des verbes symétriques à trois arguments dont le schéma verbal est de type  $\langle N_0$  V  $N_a$  and  $N_b \rangle$ . Le GNpréd correspondant est  $\langle N_v$  of/between  $N_a$  and  $N_b$  by  $N_0 \rangle$ . La nominalisation *association* appartient à cette classe. Cette nominalisation est également associée au GNpréd équivalent  $\langle N_v$  of  $N_a$  with  $N_b$  by  $N_0 \rangle$  qui est lié au schéma verbal  $\langle N_0$  V  $N_a$  with  $N_b \rangle$ . Dans ce dernier cas, la nominalisation a l'extension nsd3.

### **Le traitement des liens argumentaux**

Conformément à notre classification des nominalisations (Cf. Section 4.4), une grammaire formée de 89 sous-classes a été développée [Royauté et al., 2006, 2007, Godbert and Royauté, 2009]. Chaque sous-classe correspond à un patron syntaxique particulier. Lorsqu'une nominalisation possède plusieurs descriptions syntaxiques, elle apparaît dans la grammaire plusieurs

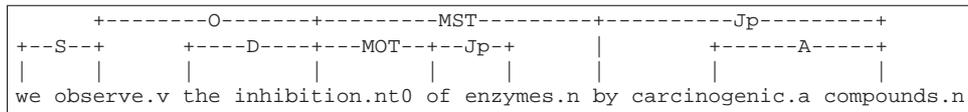


FIGURE 4.8 – Analyse d’un GNpréd

mots	formules
a the	D+
we	S+
enzymes.n compounds.n	({A-} & {D-} & {Mp+} & (O- or S+)) or ({A-} & {D-} & Jp-)
inhibition.nt0	({A-} & {D-} & ({MOT+} & {MST+}) & (S+ or O-)) or ({A-} & {D-} & Jp-)
observe.v inhibite.v	S- & 0+
big.a carcinogenic.a	A+
of	(Mp- & Jp+) or (MOT- & Jp+)
by	(Mp- & Jp+) or (MST- & Jp+)

TABLE 4.4 – Une grammaire simplifiée permettant l’analyse de GNpréd transitifs simples

fois avec des extensions différentes. Pour pouvoir faire apparaître les liens argumentaux, de nouveaux liens ont été ajoutés afin de relier les différentes prépositions qui introduisent les actants nominaux au nom prédicatif de tête. Pour cela, différents types de liens prépositionnels ont été utilisés. Ces derniers dépendent de la fonction syntaxique des actants nominaux (sujet, COD, complément prépositionnel, etc.) et de la catégorie syntaxique du nom de tête (transitif, intransitif, etc.). Ainsi, l’analyse du GNpréd de la figure 4.8 montre que la nominalisation *inhibition*, qui est dérivée d’un verbe transitif direct, possède l’extension nt0. Les prépositions *of* et *by* sont reliées à ce nom de tête, respectivement, par les liens MOT et MST. Cela signifie que la préposition *of* introduit le COD et que la préposition *by* introduit le sujet. Les autres types de liens argumentaux seront vus dans la Section 4.5.3. La table 4.4 représente une grammaire simplifiée qui a permis de produire l’analyse de la figure 4.8. Nous commentons ci-dessous les nouveaux liens et formules utilisées :

- les noms *enzymes* et *compounds* sont associés à une formule composée de deux sous-formules reliées par l’opérateur OU :
  1. la première sous-formule représente le cas où un de ces noms est le sujet d’un prédicat se trouvant à sa droite (lien S+) ou l’objet d’un prédicat qui se trouve à sa gauche (lien O-). Outre les connexions déjà connues, nous remarquons la possi-

- bilité d'une connexion prépositionnelle droite avec le lien  $M_{\mathcal{P}+}$  qui est optionnel.
2. la deuxième sous-formule représente le cas de l'insertion de ce nom dans un groupe prépositionnel. La connexion est établie par le lien  $J_{\mathcal{P}-}$  avec une préposition. Le GP *by carcinogenic compounds* de la figure 4.8 en est un exemple.
- le nom de tête *inhibition* est également associé à deux sous-formules reliées par le connecteur OU :
1. la première sous-formule est une variante de la première sous-formule des noms *enzymes* et *compounds* dont l'optionnalité du lien général  $M_{\mathcal{P}+}$  est substituée par l'optionnalité de deux liens argumentaux :  $MOT+$  et  $MST+$ . Ces derniers liens peuvent pointer respectivement sur les deux GP ayant une fonction COD et sujet dans le GNpréd, tel que le montre la figure 4.8.
  2. la seconde sous-formule est identique à la seconde sous-formule des noms *enzymes* et *compounds*.
- les prépositions *of* et *by* sont également associées à deux sous-formules reliées par l'opérateur OU :
1. Dans le cas d'un lien avec un nom de tête non prédicatif, elles se connectent avec le nom qui leur succède avec le lien  $J_{\mathcal{P}+}$  et à celui qui les précède avec le lien  $M_{\mathcal{P}-}$ .
  2. Dans le cas d'un lien avec un nom de tête prédicatif, *of* et *by* se connectent avec le nom qui est à leur droite avec  $J_{\mathcal{P}+}$  et au nom de tête, qui se trouve à leur gauche, respectivement avec les liens argumentaux  $MOT-$  (cas de la préposition *of*) et  $MST-$  (cas de la préposition *by*).

### 4.5.3 Analyse des GNpréd avec le Link Parser

Nous présentons ici plusieurs exemples d'analyses de GNpréd complexes qui appartiennent aux différentes classes que nous avons présentées (Cf. Section 4.4). Tous ces exemples sont tirés d'articles scientifiques ou de résumés de la base Medline. Quand la phrase est trop longue, elle est réduite à la partie qui nous intéresse.

#### GNpréd de type $N_v$ of $N_1$ by $N_0$

Dans l'exemple de la figure 4.9, la première nominalisation *inhibition* appartient à la super-classe 4.

Cette nominalisation possède l'extension *nt0* pour signifier qu'elle est dérivée d'un verbe tran-

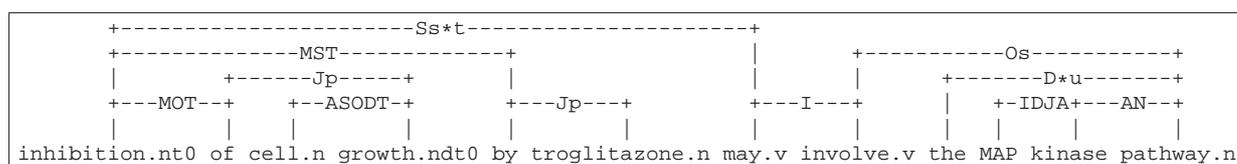


FIGURE 4.9 – Analyse d'un GNpréd ayant la forme  $N_v$  of  $N_1$  by  $N_0$

sitif direct. L'analyse montre que, comme pour l'exemple simplifié de la figure 4.8, le nom de tête *inhibition* est lié aux prépositions *of* et *by* avec, respectivement, les liens MOT et MST. Ainsi que nous l'avons vu, ces liens signifient que ces prépositions introduisent respectivement le COD *cell growth* et le sujet *troglitazone*. Le lien ASODT qui lie l'argument *cell* à la nominalisation *growth* signifie qu'il existe une ambiguïté concernant le rôle de cet argument, c'est-à-dire qu'il n'est pas possible de déterminer s'il est sujet ou objet de *growth*. Les liens *Ss\*t* et *Os* signifient, respectivement, que *inhibition* est le nom de tête du sujet du modal *may* et que le nom *pathway* est le nom de tête de l'objet du verbe *involve*. Ce verbe est considéré comme un infinitif dépendant du modal *may*. Pour cette raison, les deux verbes sont reliés entre eux par le lien I. Le lien AN connecte le nom *pathway* à son modifieur nominal *kinase*. L'analyseur considère *MAP kinase* comme étant une expression complexe de type terme. Il lie donc ces deux membres avec le lien spécifique IDJA.

#### GNpréd de type $N_v$ of $N_0$ prep $N_1$

Dans le premier GNpréd de la figure 4.10 (exemple 2-a), l'analyseur sélectionne la nominalisation *response*, appartenant à la super-classe6 avec l'extension *ni2* car il considère que le GNpréd est associé à un schéma transitif indirect construit avec la préposition *to* ( $N_v$  of  $N_0$  to  $N_1$ ). Dans l'analyse, le lien MSI identifie le sujet *erythropoietin gene* avec la préposition *of*, alors que le lien MCITO marque le complément prépositionnel (*hypoxia*) introduit par la préposition *to*, hérité du verbe. Le lien Pv connecte les différentes formes de *be* aux participes passés. Comme pour toutes les formes à deux verbes (*may involve* de l'exemple précédent ou *is mediated* de cet exemple), l'analyseur connecte le premier verbe (ici *is*) au nom de tête en position sujet (ici *response*) avec le lien Ss et le second verbe soit à un objet (ici *involve*) soit (ici *mediated*) à la préposition qui introduit son complément éventuel (ici la préposition *by*) par le lien MVP.

Dans le second exemple de la même figure, *ni4* est l'extension du nom de tête *heterodimerization*. Cette dernière identifie un GNpréd à schéma transitif indirect se construisant cette fois avec la préposition *with* (super-classe 6). Dans cette analyse également, le lien MSI identifie

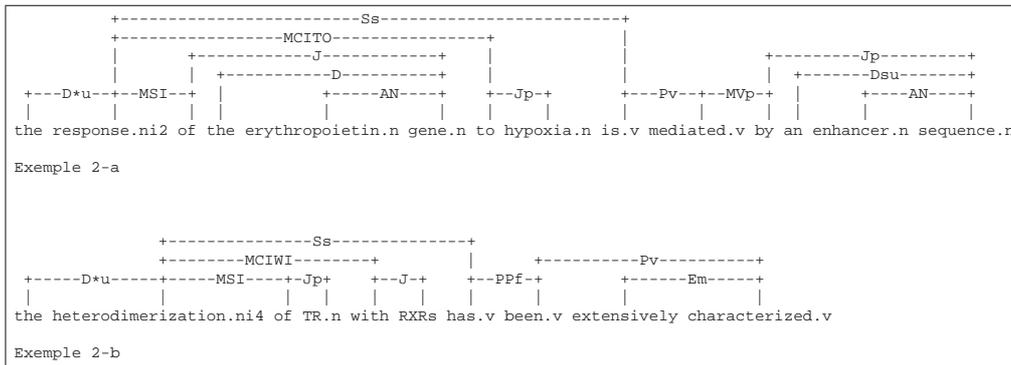


FIGURE 4.10 – Analyse de GNpréd de la forme  $N_v$  of  $N_0$  prep  $N_1$

le sujet avec la préposition *of*, alors que le lien MCIWI marque le complément prépositionnel introduit par *with*. Les différentes formes de *have* peuvent se connecter aux participes passés en utilisant le lien PPf. Le lien EM sert à connecter les adverbes qui sont modificateurs de verbes de ces derniers.

### GNpréd de type $N_v$ of $N_1$ prep $N_2$ by $N_0$

Le nom de tête *release*, qui appartient à la super-classe 4 du premier GNpréd (Figure 4.11) a ndt3 comme extension. Il est associé à un schéma à deux compléments de type  $\langle N_v$  of  $N_1$  from  $N_2$  by  $N_0 \rangle$ . Dans cette analyse on trouve trois types de liens argumentaux : (i) le lien MODT identifie le COD avec la préposition *of*, (ii) le lien MCDTFR identifie le complément prépositionnel avec la préposition *from* et (iii) le lien MSDT identifie le sujet avec la préposition *by*.

Dans le deuxième exemple de la même figure, le nom de tête *protection*, qui appartient à la même super-classe, a pour extension ndt5 et permet d'analyser des GNpréd de type  $\langle N_v$  of  $N_1$  against  $N_2$  by  $N_0 \rangle$ . L'analyse de cet exemple montre également que le lien MODT identifie le COD avec la préposition *of* et que le lien MCDTAG identifie le complément prépositionnel avec la préposition *against*. On remarque dans ce GNpréd que l'actant sujet est effacé.

### Les GNpréd dont sujet et complément(s) sont permutable

L'exemple 4-a de la figure 4.12 montre que le nom de tête *coimmunoprecipitation* (super-classe 6) possède l'extension ns1 car il est associé à un verbe symétrique dont le GNpréd

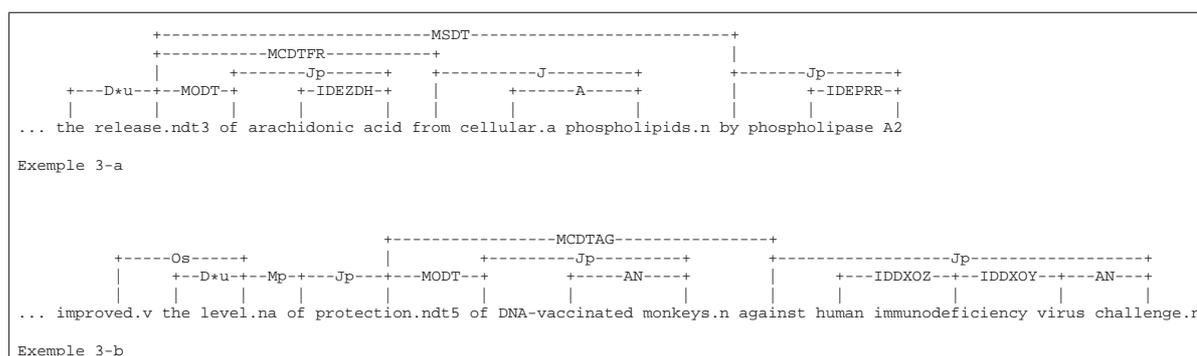


FIGURE 4.11 – Analyse de GNpréd de la forme  $N_v$  of  $N_1$  prep  $N_2$  by  $N_0$

correspondant est de type  $\langle N_v$  of  $N_a$  with  $N_b \rangle$ . Comme le montre cet exemple, les nouveaux liens MAS et MBSWI ont été utilisés pour identifier les co-agents  $N_a$  et  $N_b$ . Alors que le premier co-agent est en position sujet, le second est introduit par la préposition *with*. Le lien N sert à connecter l’adverbe *not* au verbe qui le précède. Dans l’exemple 4-b, la même nominalisation *coimmunoprecipitation* possède, dans ce cas, l’extension ns2 car elle est associée au schéma nominal équivalent  $\langle N_v$  of  $N_a$  and  $N_b \rangle$ . Le lien OFDIS, pointant sur la préposition *of*, distribue les arguments permutable  $N_a$  et  $N_b$ , marqués par les liens JAS et JBS, autour de la coordination *and* (lien AND). Le nom de tête *association*, qui appartient à la super-classe 2 de l’exemple 4-c possède l’extension nsd3 pour signifier qu’il est associé à une forme symétrique à trois arguments de type  $\langle N_v$  of  $N_a$  with  $N_b$  by  $N_0 \rangle$ . Comme le montre l’analyse correspondante, les liens MAS et MBDSWI ont été créés pour identifier les compléments permutable  $N_a$  et  $N_b$ . Cette même nominalisation possède, dans l’exemple 4-d, l’extension nsd2 car elle correspond à un autre schéma symétrique à trois arguments équivalent de type  $\langle N_v$  of  $N_a$  and  $N_b$  by  $N_0 \rangle$ . Les co-agents  $N_a$  et  $N_b$  (identifiés par les liens JAS et JBS) sont distribués par le lien OFDIS de la même façon que dans l’exemple précédent.

#### 4.5.4 Pertinence du lexique et des analyses

Les exemples d’analyses que nous avons présentés montrent l’intérêt et la pertinence de notre travail concernant les nominalisations ainsi que la capacité de notre lexique à permettre de marquer correctement les structures argumentales nominales. Précisons que les phrases analysées sont toutes issues en totalité ou en partie de travaux en bio-médecine. Elles ont toutes été soumises à PredXtract, qui à chaque fois a retenu la meilleure analyse. La pertinence de notre travail est également attestée par une évaluation récente de la plate-forme PredXtract [Godbert

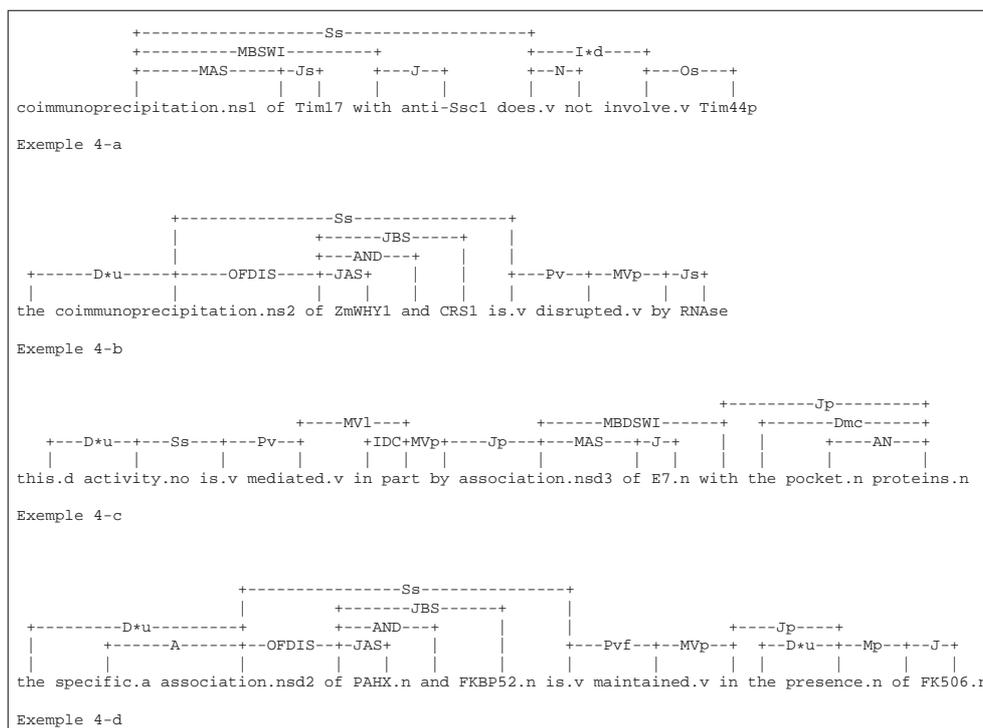


FIGURE 4.12 – Analyse de GNpred dans lesquels le sujet/compléments sont permutable

and Royauté, 2009]. Celle-ci a été réalisée à partir de 300 phrases sélectionnées aléatoirement dans un corpus de 3 500 phrases. Concernant les structures argumentales nominales, les valeurs de F-mesure obtenues sont de 0.78 pour l’identification des structures argumentales complètes. Ce résultat, plutôt bon, montre que le lexique PredicateDB a la capacité de permettre l’analyse d’une grande variété de GNpred.

## 4.6 Conclusion

Nous avons vu dans ce chapitre que le groupe nominal prédicatif possède la même structure distributionnelle qu’un groupe nominal. Il se compose d’un nom de tête prédicatif autour duquel se regroupent des arguments nominaux auxquels on peut associer des fonctions syntaxiques bien définies (sujet, complément d’objet direct, etc.). Les rôles des arguments du GNpred sont difficilement prédictibles quand ils sont placés à gauche du nom. Par contre, leur position comme post-modifieur rend leurs rôles plus facilement déductibles. En effet, en connaissant la catégorie verbale auquel le prédicat nominal dérive, on peut considérer les prépositions, qui

sont les introductrices des groupes nominaux, comme marqueurs de ces arguments.

En se basant sur des faits linguistiques connus, on a pu mettre en évidence deux grandes catégories de nominalisations : celles pour lesquelles la préposition *of* introduit un argument COD et la préposition *by* est le marqueur privilégié du sujet et celles pour lesquelles cette même préposition *of* introduit un argument sujet. Le premier type de nominalisations est associé aux schémas verbaux transitifs, c'est-à-dire ceux qui admettent un COD. Le second type est associé aux schémas verbaux intransitifs qui n'admettent pas de COD mais qui peuvent admettre des éventuels compléments prépositionnels.

À partir des données du Specialist Lexicon, nous avons développé une plate-forme (PredicateDB), organisée autour d'une base de données et de différents outils permettant sa manipulation. Cette plate-forme nous a assisté dans la confirmation du rôle que jouent les prépositions dans les différents types de GNpréd en mettant en évidence les liens qui existent entre certains schémas verbaux et les GNpréd qui leur correspondent. Le choix des schémas verbaux est basé sur une typologie qui exclut les compléments phrastiques et qui porte uniquement sur les compléments nominaux. PredicateDB nous a également permis de mettre en évidence certains liens plus complexes qui impliquent des prédicats symétriques.

PredicateDB nous a également permis de créer un lexique de nominalisations. Pour cela, nous avons classé les nominalisations en utilisant des heuristiques qui prennent en considération le comportement syntaxique de ces nominalisations afin de réduire au maximum les ambiguïtés liées au rôle de la préposition *of*.

Le lexique produit a été utilisé dans le développement d'une grammaire de liens qui a été intégrée dans la grammaire du Link Parser dans le but de rendre possible l'analyse syntaxique des GNpréd dans le domaine biologique [Royauté et al., 2006, 2007, Godbert and Royauté, 2009]. Pour cela, nous avons adapté notre lexique pour qu'il soit utilisable par le Link Parser en attribuant à chaque entrée (nominalisation) du lexique une extension qui marque le type du schéma syntaxique qui lui est associée ainsi que les prépositions qui font partie de ses entrées lexicales. Pour pouvoir identifier les liens argumentaux nominaux (sujet, COD, complément prépositionnel), la grammaire native du Link Parser a été enrichie avec de nouveaux liens. Les différentes analyses que nous présentons ainsi que l'évaluation réalisée récemment montrent l'intérêt de notre travail sur les nominalisations ainsi que la pertinence de notre lexique pour marquer correctement les structures argumentales nominales.

## Chapitre 5

# Conclusion

Notre travail a porté sur les structures prédicatives nominales pour l'analyse des textes, tout particulièrement les textes scientifiques. Nous nous sommes attaché à préciser les relations qui lient les structures syntaxiques des schémas verbaux à celles des groupes nominaux prédicatifs qui leur sont associés et avons montré comment ces relations influent sur le comportement syntaxique des arguments nominaux. Dans une perspective d'extraction d'informations, nous nous sommes intéressé, dans un premier temps, aux structures distributionnelles et fonctionnelles du groupe nominal prédicatif (GNpréd) ainsi qu'aux différentes formes équivalentes qu'un GNpréd peut avoir. Cette étude préalable nous a permis de réaliser la plate-forme PredicateDB, une base de données et un ensemble d'outils, permettant d'accéder aux informations syntaxiques des verbes et des nominalisations du Specialist Lexicon [Browne et al., 2000] et de déduire la structure argumentale des GNpréd.

### **Relations entre structures prédicatives verbales et nominales**

Ainsi que nous l'avons vu, un GNpréd peut admettre plusieurs formes équivalentes dans lesquelles les différents arguments nominaux peuvent être dans différentes positions : déterminant possessif, prémodifieur sous la forme d'un nom et en position postmodifieur introduits par des prépositions. Un GNpréd se caractérise aussi par le fait que tous ses arguments peuvent être effacés. Notre travail ne s'est intéressé qu'au cas où les arguments nominaux sont en position postmodifieur.

Le comportement syntaxique des GNpréd dépend fortement de la classe grammaticale du prédicat verbal à partir duquel la nominalisation a été dérivée et de son schéma de complémentation. Les prédicats verbaux peuvent être répartis en deux grandes classes : les prédicats transitifs qui admettent au moins un COD, de la forme

$\langle N_0 V N_1 [\text{Prep } N_2 \dots \text{Prep } N_n] \rangle$  et les prédicats qui n'admettent pas de COD, de la forme  $\langle N_0 V [\text{Prep } N_1 \dots \text{Prep } N_n] \rangle$ . Nous avons vu que la particularité de la première classe est qu'elle sélectionne son argument COD, dans la plupart des cas, avec la préposition *of*. L'argument sujet, lui, introduit par la préposition *by* est hérité du passif de la forme verbale. Si la construction verbale admet aussi des compléments prépositionnels, nous les retrouvons dans la structure du GNpréd avec les mêmes prépositions. Les GNpréd associés aux schémas verbaux de la deuxième classe admettent un argument sujet qui cette fois-ci est introduit, dans la majorité des cas, par la préposition *of*. Comme pour la première classe, les éventuels compléments prépositionnels du verbe dérivé se retrouvent à l'identique dans la structure du GNpréd. On remarque que la préposition *of* prise isolément est très ambiguë. Cependant, cette ambiguïté disparaît dans les GNpréd saturés, quand il est possible d'associer la nominalisation à l'une des deux classes.

### **Généralisation des constructions verbales aux constructions nominales**

Les propriétés que nous avons mises en évidence ont été le point de départ de la création d'une plate-forme, PredicateDB, dont le but est l'obtention d'un lexique de nominalisations où chaque argument, dans le GNpréd est marqué par des prépositions spécifiques. Notre objectif en créant PredicateDB, une plate-forme réalisée autour d'une base de données et des outils permettant de la manipuler, était d'avoir un environnement permettant d'établir des corrélations entre les structures verbales et nominales du SL. Afin de prédire la structure argumentale des GNpréd et leurs marqueurs, nous avons procédé de la façon suivante : (i) détermination de l'ensemble des classes verbales du SL (rappelons que très souvent, un verbe est associé à plusieurs classes verbales), (ii) mise en correspondance du schéma nominal associé à chacune de ces classes en utilisant la macro-classification permettant de distinguer les GNpréd introduisant leur sujet ou leur objet avec *of*, (iii) regroupement de ces structures de GNpréd en super-classes à partir de leurs nominalisations sur la base de GNpréd qui sont non ambigus par rapport au rôle de la préposition *of*. Il est ainsi possible de générer un lexique dont les entrées sont les nominalisations et leurs différentes propriétés, c'est-à-dire, pour chaque nominalisation : les formes saturées liées à chaque emploi verbal et les cas d'indécidabilité de rattachement à un emploi verbal du fait du rôle ambigu de la préposition *of* en tant que marqueur pour les formes non saturées.

PredicateDB permet également d'affiner semi-automatiquement la description syntaxique à partir d'hypothèses linguistiques. Nous avons utilisé ces possibilités de manipulation pour identifier des constructions nominales liées à des verbes symétriques. Rappelons qu'un grand nombre d'entre elles, comme nous l'avons vu, se construisent pour leur verbe sous la

forme :  $\langle N_a V \text{ with } N_b \rangle$  où les arguments  $N_a$  et  $N_b$  ont la propriété d'être permutable. Avec PredicateDB, nous avons extrait ces constructions que nous avons pu croiser avec des emplois nominaux avec *with* et *between*. Nous avons de cette façon, semi-automatiquement, mis en évidence les nominalisations qui partagent ces propriétés.

À partir de procédures automatiques et semi-automatiques, nous avons pu ainsi créer un lexique de nominalisations que nous avons utilisé dans le développement d'une grammaire de liens pour le Link Parser dans le but de réaliser l'analyse syntaxique des GNpréd appartenant au domaine de la biologie [Royauté et al., 2006, 2007, Godbert and Royauté, 2009]. Pour cela, nous avons enrichi la grammaire native du Link Parser avec de nouveaux liens qui permettent de marquer d'une manière pertinente les arguments nominaux. L'évaluation réalisée atteste de la pertinence et de l'intérêt de notre lexique pour marquer correctement les structures argumentales nominales.

## Discussion et Perspectives

Nous avons utilisé PredicateDB avec un seul lexique, le Specialist Lexicon, qui couvre un grand nombre de termes de la biologie et des sciences humaines et sociales. Cependant, ce lexique nous donne principalement des informations distributionnelles sur les arguments. Nous montrons maintenant comment il est possible d'aller plus loin dans cette direction afin d'affiner et étendre la ressource créée.

**Nominalisations à compléments phrastiques :** Nous avons vu (Cf. Section 4.4.5) qu'il y avait un intérêt à traiter les nominalisations qui admettent des compléments phrastiques à partir d'un sous-ensemble de prédicats dont les compléments sont de type : infinitif, participe présent, complétive et formes introduites par des mots en WH (*whether, when, who, etc.*) Une étude linguistique approfondie devrait permettre de vérifier au cas par cas les possibilités d'insertion de ce type de compléments dans la structure du GNpréd. Rappelons que ces informations concernant les verbes sont disponibles dans la base PredicateDB.

**Nominalisations rattachées à des verbes à plusieurs emplois verbaux :** Un grand nombre de verbes du Specialist Lexicon acceptent deux emplois : intransitif et transitif. Dans le cas d'une nominalisation non saturée avec *of*, nous avons admis que ces deux emplois n'étaient pas reliés et par conséquent, il n'était pas possible de décider si l'argument marqué par *of*, dans le cas d'une forme non saturée était un sujet ou un objet. Cependant, après examen des données, le cas de deux emplois verbaux non reliés, nous semble une possibilité, qui si elle se présente devrait être rare et concerner des sens différents. Il est souhaitable dans ce cas de faire

l'hypothèse que ces deux schémas soient liés. En examinant un sous-ensemble de ces prédicats, nous pouvons envisager les possibilités suivantes :

- le premier cas concerne les formes transitives pour lesquelles il est possible d'effacer le COD. Il s'agit de cas peu nombreux comme avec *broadcast*, verbe et nominalisation. Dans ce cas, la forme sans COD est une forme réduite et le GNpréd se construira, ainsi que nous l'avons vu, avec *of* et *by* (exemples 5.1a et 5.1b). La forme non saturée avec *of* introduira donc toujours un COD.

(5.1) (a) *BBC radio broadcasts [sports events]*

(la radio BBC diffuse [des événements sportifs])

(b) *the broadcast of sports events [by BBC radio]*

(la diffusion d'événements sportifs [par la radio BBC])

- le second cas concerne la voix moyenne, illustrée par le verbe *accumulate* (exemples 5.2a et 5.2b). La voix moyenne telle qu'on peut le constater dans la phrase 5.2b est apparentée à un passif sans agent et peut être représentée de la façon suivante :  $\langle N_0 V N_1 \rangle$  et  $\langle N_1 V \rangle$ . Dans la forme nominale réduite *the accumulation of the energy*, la préposition *of* est bien introductrice de  $N_1$  et non pas du  $N_0$ .

(5.2) (a) *the solar pannel accumulates the energy*

(le panneau solaire accumule l'énergie)

(b) *the energy accumulates in the body*

(l'énergie s'accumule dans le corps)

- le troisième cas, beaucoup plus rare, concerne les emplois réfléchis, qui le plus souvent, ne sont pas reliés à une forme transitive. Il y a donc changement de sens ou glissement de sens entre les deux formes. Comme par exemple avec *reproduce* (exemples 5.3a et 5.3b).

(5.3) (a) *the insects reproduce rapidly*

(les insectes se reproduisent rapidement)

(b) *test program reproduces the problems*

(le programme de test reproduit les problèmes)

Le premier sens concerne la propriété des insectes à se reproduire, alors que le second,

d'un point de vue générique, est la capacité à réitérer un processus capable d'engendrer un objet concret ou abstrait qui peut éventuellement être considéré comme un clone de l'objet initial.

L'examen des structures nominales montre que pour deux d'entre-elles (la forme transitive avec réduction du  $N_1$  ( $N_0$  V [ $N_1$ ]) et la forme transitive liée à la voix moyenne  $N_1$  V), il n'y a pas de changement concernant le rôle de la préposition *of* en cas de GNpréd non saturé. Dans les deux cas, cette préposition est introductrice de  $N_1$  ( $N_v$  of  $N_1$ ). Concernant les GNpréd de notre troisième cas, qui sont liés à une forme transitive et une forme réfléchie, seule la catégorie sémantique de son argument pourrait décider du rôle argumental de la préposition *of* dans le cas d'un GNpréd à un seul argument. Remarquons que ces deux premiers cas sont les plus nombreux et le troisième est beaucoup plus rare. Grâce à notre plate-forme, l'examen manuel serait grandement facilité car cela ne nécessiterait qu'un parcours où seul un petit nombre de prédicats requièrerait un examen approfondi. Cet exemple simple ainsi que tous ceux que nous avons présentés dans notre thèse montrent l'intérêt de l'utilisation de PredicateDB en son état à partir d'un raisonnement linguistique. D'autres explorations de configurations linguistiques plus complexes peuvent être envisagées avec profit.

**Nominalisations dérivées d'adjectifs :** Outre les nominalisations verbales, le SL recense également les nominalisations d'adjectifs. Il nous semble important de traiter ces nominalisations comme celles des verbes.

- (5.4) (a) *p53 gene is aberrant*  
(le gène p53 est aberrant)  
(b) *aberration of p53*  
(l'aberration du gène p53)

- (5.5) (a) *protein is accessible to enzymatic attack*  
(la protéine est accessible à une attaque enzymatique)  
(b) *accessibility of protein to enzymatic attack*  
(l'accessibilité de la protéine à une attaque enzymatique)

Ces deux exemples montrent deux emplois possibles d'adjectifs dérivés en noms. L'exemple 5.4 porte sur les emplois les plus fréquents des adjectifs et de leurs nominalisations où le seul argument est le sujet. L'exemple 5.5, moins fréquent, illustre le cas où l'adjectif et sa nominali-

sation ont deux arguments : un argument sujet et un argument complément, ici introduit par la préposition *to*. Nous retrouvons ces informations dans le SL, ce qui rend possible une extension de ce travail aux nominalisations d'adjectifs, bien que pour le moment nous n'avons pas intégré ces informations dans la base PredicateDB.

**Couplage avec d'autres ressources :** l'utilisation d'autres ressources linguistiques tels que NOMLEX, VerbNet, FrameNet devrait nous permettre de valider les résultats obtenus avec PredicateDB. On pourrait également les utiliser pour rechercher automatiquement ou semi-automatiquement les propriétés qui nous intéressent pour l'enrichissement. Cependant, l'intégration dans une base de données nécessiterait un important travail conceptuel qui n'est pas à notre portée pour le moment. Par contre, il serait envisageable d'intégrer certains segments en fonction des propriétés qui nous font défaut. Par exemple, VerbNet marque les verbes qui admettent la voix moyenne (*Cf.* Section précédente) avec l'étiquette *middle construction*. On peut utiliser ces informations pour extraire tous les verbes qui admettent ce type de construction et en exploitant PredicateDB, les croiser avec les informations fournies par le Specialist Lexicon pour confirmer les schémas qui leur sont associés. VerbNet contient 945 verbes qui possèdent la propriété *middle construction*. Parmi ces 945 verbes, 392 possèdent des nominalisations dans le SL. Par exemple le verbe *accelerate* (accélérer), qui est défini dans VerbNet comme étant un verbe qui admet la construction moyenne, est présent dans le SL avec la nominalisation *acceleration*. Ce verbe est associé à deux schémas : (i)  $\langle N_0 V N_1 \rangle$  (exemple 5.6a) et (ii)  $\langle N_1 V \rangle$  (exemple 5.6c). Dans le premier exemple, *deforestation* joue le rôle de l'argument  $N_1$  et ce même argument, dans le second exemple, conserve sa fonction syntaxique mais occupe la place du sujet dans une forme réfléchie. On remarque que ces deux schémas sont associés au même GNpréd :  $\langle N_v \text{ of } N_1 [\text{by } N_0] \rangle$ , où  $\langle \text{by } N_0 \rangle$  est optionnel et dans lequel la préposition *of* introduit le COD (exemples 5.6b et 5.6d).

L'intégration de propriétés sémantiques afin de donner une représentation abstraite du sens d'un prédicat et de ces arguments nous paraît plus délicate. En effet, les propriétés qui ont été spécifiées dans VerbNet ou FrameNet reposent sur un niveau de langue général qui n'intègre pas les spécificités d'un sous-langage. Cependant, la classification de Levin enrichie dans VerbNet pour ces verbes pourrait nous être utile car un verbe qui appartient à une langue de spécialité et qui est également utilisé dans la langue générale, présente le plus souvent un invariant sémantique qui lui permet d'être utilisé de façon identique dans les deux niveaux de langue.

(5.6) (a) *return of refugees accelerates deforestation*

(le retour des réfugiés a accéléré la déforestation)

(b) *the acceleration of deforestation by the return of refugees*

- (l'accélération de la déforestation par le retour des réfugiés)
- (c) *deforestation accelerates*  
(la déforestation s'accélère)
- (d) *the acceleration of deforestation*  
(l'accélération de la déforestation)

**Caractérisation des arguments en position modifieur ou possessif :** une difficulté importante dans l'analyse des GNpréd est de pouvoir préciser le rôle d'un argument quand celui-ci est en position gauche ou déterminant possessif. Un travail intéressant reste à réaliser afin de vérifier si les informations de classification des verbes de VerbNet pourraient être utilement croisées avec celles de NOMLEX. Nous pouvons espérer ainsi mettre en évidence des régularités caractérisant ces arguments en position gauche.

Nous pensons que l'extraction de l'information dans le domaine de la biologie est un domaine de recherche qui est appelé à se développer et permettra, dans le contexte de l'Internet et de l'importance des ressources électroniques, de mieux maîtriser les flux d'informations. Dans ce domaine, comme d'une façon générale dans la langue scientifique, les nominalisations sont très nombreuses. Or, comme nous l'avons montré durant cette thèse, la mise en évidence des structures argumentales des prédicats nominaux pose des problèmes délicats d'analyse automatique, bien plus délicats que pour les prédicats verbaux. Notre base PredicateDB, ainsi que les outils que nous avons développés pour son utilisation, permettent d'acquérir les données lexicales nécessaires à l'analyse automatique des GNpréd. Une analyse correcte de ce type d'objet linguistique, ainsi que nous l'avons montré (Cf. Section 4.5) en utilisant une grammaire de dépendances, illustre la pertinence de notre démarche. De plus, une telle analyse pour l'extraction d'informations permet d'exhiber une information qui aurait été cachée avec des moyens plus traditionnels d'analyse syntaxique automatique.

# Bibliographie

- A. Alexiadou. *Functional Structure in Nominals : Nominalization, and Ergativity*. Amsterdam/Philadelphia : John Benjamins, 2001.
- E. L. Antworth. *PC-KIMMO : a two-level processor for morphological analysis*. Occasional Publications in Academic Computing No. 16. Dallas, TX : Summer Institute of Linguistics, 1990. ISBN ISBN 0-88312-639-7.
- C. Aone and D. McKee. Acquiring predicate-argument mapping information from multilingual texts. pages 191–202, 1996.
- M. Aronoff. *Word Formation in Generative Grammar*. Cambridge : MA : MIT Press, 1976.
- S. Bangalore and A. K. Joshi. Supertagging : an approach to almost parsing. *Comput. Linguist.*, 25(2) :237–265, 1999. ISSN 0891-2017.
- B. Boguraev and J. Pustejovsky. Issues in text-based lexicon acquisition. pages 3–17, 1996.
- M. R. Brent. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 209–214, Morristown, NJ, USA, 1991. Association for Computational Linguistics. doi : <http://dx.doi.org/10.3115/981344.981371>.
- M. R. Brent. From grammar to lexicon : unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2) :243–262, 1993. ISSN 0891-2017.
- T. Briscoe and J. Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the fifth conference on Applied natural language processing*, pages 356–363, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.
- A. C. Browne, A. T. McCray, and S. Srinivasan. The SPECIALIST LEXICON technical report. *Lister Hill National Center for Biomedical Communications, National Library of Medicine, USA.*, 2000.

- L. Burnard. *Users Guide for the British National Corpus*. British National Corpus Consortium, oxford university computing service edition, 1995.
- D. A. Cruse. *Lexical Semantics*. Cambridge University Press, Cambridge, UK, 1986.
- D. Crystal. *A Dictionary of Linguistics & Phonetics*, volume Fifth edition. Blackwell Publishing Ltd, 108 Cowley Road, Oxford OX4 1JF, UK, 2003. ISBN 0-631-22663-X.
- H. T. Dang, K. Kipper, M. Palmer, and J. Rosenzweig. Investigating regular sense extensions based on intersective levin classes. In *Proceedings of the 17th international conference on Computational linguistics*, pages 293-299, Morristown, NJ, USA, 1998. Association for Computational Linguistics. doi : <http://dx.doi.org/10.3115/980845.980893>.
- Z. Dong. HowNet Chinese-English Conceptual Database. In *Technical Report Online Software Database, Released at ACL*. <http://www.keenage.com>, 2000.
- C. Doran, D. Egedi, B. Hockey, B. Srinivas, and M. Zaidel. XTAG system — a wide coverage grammar for English. In *COLING-94*, pages 922-928, 1994.
- B. J. Dorr. Unitran : An interlingual machine translation system. Technical report, Cambridge, MA, USA, 1987.
- B. J. Dorr. Machine translation divergences : a formal description and proposed solution. *Comput. Linguist.*, 20(4) :597-633, 1994. ISSN 0891-2017.
- B. J. Dorr. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(4) :271-322, 1997.
- B. J. Dorr and D. Jones. Role of word sense disambiguation in lexical acquisition : predicting semantics from syntactic cues. In *Proceedings of the 16th conference on Computational linguistics*, pages 322-327, Morristown, NJ, USA, 1996. Association for Computational Linguistics. doi : <http://dx.doi.org/10.3115/992628.992685>.
- B. J. Dorr, M. A. Martí, and I. Castellón. Spanish EuroWordNet and LCS-Based Interlingual MT. In *Proceedings of the Workshop on Interlinguas in MT, MT Summit, San Diego, CA*, pages 19-32, October 1997.
- J. Dubois, J. P. Mével, G. Chauveau, S. Hudelot, C. Sobotka-Kannas, and D. Morel. *dictionnaire de la langue française LEXIS*. Librairie Larousse, 17 rue du Montparnasse, 75298 Paris CEDEX 06, 1989. ISBN 2-03-320211-9.

- J. Dubois, M. Giacomo, L. Guespin, C. Marcellesi, J. B. Marcellesi, and J. P. Mével. *Dictionnaire de Linguistique*. Larousse, 17, Rue de Montparnasse, 75298 Paris Cedex 06, 1991. ISBN 2-03-340308-4.
- C. Fellbaum. *WordNet : An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998. ISBN 026206197X.
- C. Fellbaum and G. A. Miller. Morphosemantic Links in WordNet. *Traitement automatique de la langue*, 44(2) :69-80, 2003.
- C. Fellbaum, A. Osherson, and P. E. Clark. Putting Semantics into WordNet's "Morphosemantic" Links. *Proceedings of the Third Language and Technology Conference, Poznan (Poland)*, 2007.
- J. C. Fillmore. Frame semantics. *In Linguistics in the Morning Calm. Seoul : Hanshin.*, pages 111-137, 1982.
- J. C. Fillmore and B. T. S. Atkins. Towards a frame-based lexicon : The semantics of risk and its neighbors. *In A. Lehrer and E. F. Kittay (Eds.), Frames, Fields, and Contrasts*, 75-102. Hillsdale, NJ : Erlbaum., 1992.
- W. N. Francis and H. Kucera. *Frequency Analysis of English Usage : Lexicon and Grammar*. Houghton Mifflin : Boston, Massachusetts, 1982.
- E. Godbert and J. Royauté. Exploring predicate-arguments structures in texts to relate biological entities. *Actes de la 8ème conférence internationale Terminologie et Intelligence Artificielle (TIA-2009), Atelier "Acquisition et modélisation de relations sémantiques"*, Toulouse (France), 2009.
- R. Green and B. J. Dorr. Inducing a semantic frame lexicon from wordnet data. *Workshop on Text Meaning and Interpretation, 42nd Annual Meeting of the Association of Computational Linguistics.*, 2004.
- R. Green, L. Pearl, B. J. Dorr, and P. Resnik. Mapping WordNet Senses to a Lexical Database of Verbs. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France*, pages 244-251, 2001.
- R. Green, B. J. Dorr, and P. Resnik. Inducing frame semantic verb classes from wordnet and ldoce. *42nd Annual Meeting of the Association of Computational Linguistics.*, 2004.

- G. Grefenstette and S. Teufel. Corpus-based method for automatic identification of support verbs for nominalizations. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 98–103, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. doi : <http://dx.doi.org.gate6.inist.fr/10.3115/976973.976988>.
- J. Grimshaw and R. Jackendoff. Brandeis verb lexicon. electronic database funded by national science foundation grant nsf ist-81-20403 awarded to brandeis university. 1981.
- R. Grishman, C. Macleod, and A. Meyers. Complex Syntax : Building a Computational Lexicon. *International Conference On Computational Linguistics, Proceedings of the 15th conference on Computational linguistics. Kyoto, Japan, 1, 1994.*
- M. Gross. Lexicon-grammar : the representation of compound words. *International Conference On Computational Linguistics. Proceedings of the 11th conference on Computational linguistics*, 1986.
- O. Gurevich, R. S. Crouch, T. H. King, and V. de Paiva. Deverbal nouns in knowledge representation. *Proceedings of the 19th International Florida AI Research Society Conference (FLAIRS '06)*, pages 670–675, 2006.
- N. Habash. Generation-Heavy Hybrid Machine Translation. In *In Proceedings of the International Natural Language Generation Conference (INLG'02) Student Session*, 2002.
- N. Habash and B. Dorr. A categorial variation database for English. In *NAACL '03 : Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 17–23, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi : <http://dx.doi.org/10.3115/1073445.1073458>.
- N. Habash, B. Dorr, and D. Traum. Efficient Language Generation From Lexical Conceptual Structures. *Machine Translation*, (17), 2002.
- R. Jackendoff. *Semantics and Cognition*. The MIT Press, Cambridge, MA, 1983.
- R. S. Jackendoff. *Semantic Structures*. Cambridge : MIT Press, 1990.
- C. Johnson, C. Fillmore, M. Petrucci, C. Baker, M. Ellsworth, J. Ruppenhofer, and E. Wood. FrameNet : Theory and Practice. <http://framenet.icsi.berkeley.edu/>, 2002.
- D. A. Jones, R. C. Berwick, F. Cho, Z. Khan, K. Kohl, N. Nomura, A. Radhakrishnan, U. Sauerland, and B. Ulicny. Verb classes and alternations in bangla, german, english, and korean. Technical report, Cambridge, MA, USA, 1994.

- A. K. Joshi. How much context-sensitivity is necessary for characterizing structural descriptions - tree adjoining grammars. In : D. Dowty, L. Karttunen and A. Zwicky, Editors, *Natural Language Parsing*, Cambridge University Press, Cambridge, pages 206–250, 1987.
- A. K. Joshi, L. S. Levy, and M. Takahashi. Tree Adjunct Grammars. *Journal of Computer and System Sciences*, 10(1) :136–163, 1975.
- K. Kipper, H. T. Dang, and M. Palmer. Class-based construction of a verb lexicon. *Source Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence table of contents*, pages 691–696, 2000.
- J. Klavans and M.-Y. Kan. Role of verbs in document analysis. In *Proceedings of the 17th international conference on Computational linguistics*, pages 680–686, Morristown, NJ, USA, 1998. Association for Computational Linguistics.
- J. L. Klavans and M. Chodorow. Degrees of stativity : the lexical representation of verb aspect. In *Proceedings of the 14th conference on Computational linguistics*, pages 1126–1131. Association for Computational Linguistics, 1992.
- K. Knight and S. Luk. Building a Large Knowledge Base for Machine Translation. *Proceedings of the American Association of Artificial Intelligence Conference AAAI-94*. Seattle, WA., 1994.
- K. Koskenniemi. Two-Level Morphology : A general computational model for word-form recognition and production. In *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pages 178–181, Morristown, NJ, USA, 1984. Association for Computational Linguistics.
- K. Koskenniemi and K. W. Church. Complexity, two-level morphology and finnish. In *Proceedings of the 12th conference on Computational linguistics*, pages 335–340, Morristown, NJ, USA, 1988. Association for Computational Linguistics. ISBN 963 8431 56 3. doi : <http://dx.doi.org/10.3115/991635.991704>.
- R. Krovetz. Viewing morphology as an inference process. In *SIGIR '93 : Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202, New York, NY, USA, 1993. ACM. ISBN 0-89791-605-0.
- I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, pages 704–710, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

- M. Lapata. Acquiring lexical generalizations from corpora : a case study for diathesis alternations. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 397–404, Morristown, NJ, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3.
- M. Lapata and C. Brew. Using subcategorization to resolve verb class ambiguity. In *P. Fung and J. Zhou, editors, Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pages 266–274. Association for Computational Linguistics, 1999.
- B. Levin. English Verb Classes and Alternation : A preliminary Investigation. *The University of Chicago Press*, 1993.
- C. Macleod, R. Grishman, A. Meyers, L. Barret, and R. Reeves. Nomlex : A lexicon of nominalizations. In *Proceedings of the Eighth International Congress of the European Association for Lexicography.*, pages 187–193, 1998.
- M. M. Malik and J. Royauté. PredicateDB : A tool for assisting the creation of a lexicon-grammar of Predicative Nouns . *3rd Language and Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland*, pages 250–254, 2007.
- M. M. Malik and J. Royauté. A Predicate Database for Assisting the Design of a Lexicon-grammar of Predicative Nouns. *Lecture Notes in Artificial Intelligence LNAI, ed. Z. Vetulani and H. Uszkoreit (eds.)*, pages 312–324, 2009.
- C. D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 235–242, Morristown, NJ, USA, 1993. Association for Computational Linguistics.
- M. Marcus. The penn treebank : A revised corpus design for extracting predicate argument structure. *ARPA Human Language Technology Workshop. Princeton, NJ*, March 1994.
- M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of english : the penn treebank. *Computational Linguistics*, 19, 1993.
- M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The Penn Treebank : annotating predicate argument structure. In *HLT '94 : Proceedings of the workshop on Human Language Technology*, pages 114–119, Morristown, NJ, USA, 1994. Association for Computational Linguistics. ISBN 1-55860-357-3.

- D. McCarthy and A. Korhonen. Detecting verbal participation in diathesis alternations. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, pages 1493–1495. Association for Computational Linguistics, 1998.
- M. C. McCord. Slot grammars. *Computational Linguistics*, 6(1) :31–43, 1980. ISSN 0891-2017.
- M. C. McCord. *Slot Grammar : A system for simpler construction of practical natural language grammars*, volume 459/1990 of *Lecture Notes in Computer Science*, pages 118–145. Springer Berlin / Heidelberg, 1990.
- A. T. McCray, S. Srinivasan, and A. C. Browne. Lexical Methods for Managing variation in Biomedical Terminologies. in *the Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, pages 235–239, 1994.
- I. Mel'čuk. *Dependency Syntax : Theory and Practice*. State University of New York Press, New York, 1988.
- P. Merlo and S. Stevenson. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3) :373–408, 2001.
- A. Meyers, R. Grishman, M. Kosaka, and S. Zhao. Covering Treebanks with GLARF. pages 51–58, 2001a.
- A. Meyers, M. Kosaka, S. Sekine and, R. Grishman, and S. Zhao. "Parsing and GLARFing". 2001b.
- A. Meyers, R. Grishman, and M. Kosaka. Formal Mechanisms for Capturing Regularizations. *In Proceedings of LREC-2002*, 2002.
- A. Meyers, R. Reeves, C. Macleod, R. Szekeley, V. Zielinska, B. Young, and R. Grishman. The cross-Breeding of Dictionaries. *In proceedings of LREC-2004, Lisbon, Portugal*, 2004a.
- A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. The nombank project : An interim report. *In proceeding of HLT-EACL Workshop : Frontiers in Corpus Annotation.*, 2004b.
- G. Miller. WordNet : A lexical database. *Communication of the ACM* 38, 11, 39-41, 1995.
- G. A. Miller. WordNet : a dictionary browser. *Proceedings of the First International Conference on Information in Data, University of Waterloo, Waterloo*, 1985.
- G. A. Miller. Nouns in WordNet : a lexical inheritance system. *International Journal of Lexicography*, 3(4) :245–264, 1990.

- M. Moens and M. Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2) :15–28, 1988.
- D. A. Norman and D. E. Rumelhart. The active structural network. In D. E. Rumelhart, D. A. Norman, and LNR Research group (Eds.), *Explorations in Cognition*, pages 35–64, 1975.
- M. Nunes. Argument linking in English derived nominals. In Van Valin(ed)*Advances in role and reference grammar*, John Benjamins, pages 375–432, 1993.
- M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank : An Annotated Corpus of Semantic Roles. *MIT Press*, 31(1) :71–106, 2005.
- R. Pasero, J. Royauté, and P. sabatier. Sur la syntaxe et la sémantique des groupes nominaux à tête prédicative. 2004.
- M. F. Porter. An algorithm for suffix stripping. pages 313–316, 1997.
- P. Procter. *Longman Dictionary of Contemporary English*. Longman, London, 1978.
- R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A Comprehensive grammar of the English Language*. Longman, 1987.
- R. Reeves, C. Macleod, and A. Meyers. *Manual of NOMLEX : The Regularized Version*. Computer Science Department, New York University, 1999.
- P. Resnik. Selectional Constraints : An Information-Theoretic Model and its Computational Realization. *Cognition*, 61(1–2) :127–159, 1996.
- E. Riloff and M. Schmelzenbach. An Empirical Approach to Conceptual Case Frame Acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, 1998.
- J. Royauté, E. Godbert, and M. M. Malik. Groupes nominaux prédicatifs : Utilisation d’une grammaire de liens pour l’extraction d’informations. *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, pages 276–286, 2006.
- J. Royauté, E. Godbert, and M. M. Malik. Identifying relations between scientific objects within predicate structures. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 515–519, 2007.
- N. Sager. *Natural Language Information Processing : A Computer Grammar of English and Its Applications*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1981.

- A. Sanfilippo. *LKB encoding of lexical knowledge*. In T. Briscoe, A. Copestake, and V. de Pavia, editors, *Default Inheritance in Unification-Based Approaches to the Lexicon*, Cambridge University Press, 1992.
- Y. Schabes. *Mathematical and Computational Aspects of Lexicalized Grammars*. PhD thesis, Computer Science Department, University of Pennsylvania, 1990.
- S. Schulte im Walde. Automatic Semantic Classification of Verbs According to Their Alternation Behaviour. Master's thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1998.
- T. Sekimizu, H. Park, and J. Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. In *Proceeding of the 9th Workshop Genome Informatics*, pages 62–71. Universal Academy Press, 1998.
- E. V. Siegel. Corpus-based linguistic indicators for aspectual classification. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 112–119. Association for Computational Linguistics, 1999. ISBN 1-55860-609-3.
- D. Sleator and D. Temperley. Parsing English with a Link Grammar. *Carnegie Mellon University Computer Science technical report, CMU-CS-91-196, Carnegie Mellon University, USA.*, 1991.
- M. Stede. A generative perspective on verb alternations. *Comput. Linguist.*, 24(3) :402–430, 1998. ISSN 0891-2017.
- S. Stevenson and P. Merlo. Automatic verb classification using distributions of grammatical features. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 45–52, Morristown, NJ, USA, 1999. Association for Computational Linguistics. doi : <http://dx.doi.org/10.3115/977035.977043>.
- M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. Using predicate-argument structures for information extraction. In *ACL '03 : Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 8–15, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
- J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll. Automatic Extraction of Protein Interactions from Scientific Abstracts. In *Proceedings of the 5th Pacific Symposium on Biocomputing*, pages 541–553, 2000.

- D. Traum and N. Habash. Generation from lexical conceptual structures. In *NAACL-ANLP 2000 Workshop on Applied interlinguas*, pages 52–59, Morristown, NJ, USA, 2000. Association for Computational Linguistics.
- A. Ushioda, D. A. Evans, T. Gibson, and A. Waibel. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In I. B. Boguraev and e. J. Pustejovsky, editors, *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, pages 95–106, Columbus, OH, 1993.
- J. Xu and W. B. Croft. Corpus-based stemming using cooccurrence of word variants. *ACM Trans. Inf. Syst.*, 16(1) :61–81, 1998. ISSN 1046-8188.
- A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Finding anchor verbs for biomedical IE using predicate-argument structures. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 17, Morristown, NJ, USA, 2004. Association for Computational Linguistics. doi : <http://dx.doi.org.gate6.inist.fr/10.3115/1219044.1219061>.
- D. Zajic, B. Door, and R. Schwartz. Automatic Headline Generation for Newspaper Stories. In *In Proceedings of the ACL-2002 Workshop on Text Summarization (DUC 2002)*, pages 78–85, 2002.

## Résumé

Dans cette thèse, nous nous intéressons aux relations qui peuvent exister entre des prédicats verbaux (ex : *to regulate*) et des prédicats nominaux (ex : *regulation*) dont les structures argumentales mettent en jeu des informations communes. Nous nous livrons à une formalisation des conditions dans lesquelles se réalisent des relations d'équivalence entre les constructions verbales et nominales. La mise en évidence de l'équivalence des structures argumentales de ces deux types de constructions est fondamentale pour pouvoir réaliser, par exemple, des systèmes d'extraction automatique d'informations très performants. En se basant sur les données du lexique The Specialist Lexicon, nous proposons une prédiction raisonnable du comportement syntaxique des arguments nominaux, de différents groupes nominaux prédictifs (GNpréd), lorsqu'ils sont en position de postmodifieur. Cette étude nous a conduit à concevoir un ensemble d'algorithmes et à développer une plate-forme, PredicateDB, qui nous a permis de produire un lexique de nominalisations. Pour chaque entrée appartenant à ce lexique, nous avons caractérisé ses structures argumentales et ses réalisations dans des GNpréd dont les arguments sont marqués par des prépositions spécifiques.

**Mots clés :** Nominalisation, Groupe nominal prédictif, Rôle argumentale, Constructions nominales et verbales, Lexique, Grammaire de dépendance, Link Parser.

## Abstract

In this thesis, we focus on the relation that may exist between verbal predicates (e.g., *regulate*) and nominal predicates (e.g., *regulation*) whose argument structures involve common information. We make a formalization of the conditions in which equivalent relations between verbal and nominal constructions are carried out. Bringing out the equivalence of argument structures between these two types of constructions is fundamental for achieving, for example, very efficient Information Extraction systems. Based on data from the Specialist Lexicon, we propose a reasonable prediction of the syntactic behavior of nominal arguments, which belong to different predicate noun phrases (PNPs), when they are in postmodifier position. This study has led us to design a set of algorithms and develop a platform, PredicateDB, to produce a lexicon of nominalizations. For each entry belonging to this lexicon, we have defined its argument structures and achievements in PNPs whose arguments are marked by specific prepositions.

**Keywords :** Nominalization, Predicate noun phrase, Argument role, Nominal and verbal constructions, Lexicon, Dependency grammar, Link Parser.