HAL open science

Un modèle d'analyseur syntaxique fondé sur la modularité et la lexicalisation de ses grammaires

Núria Gala

▶ To cite this version:

Núria Gala. Un modèle d'analyseur syntaxique fondé sur la modularité et la lexicalisation de ses grammaires. Informatique [cs]. Université Paris Sud, 2003. Français. NNT : 2003PA112040 . tel-03988667

HAL Id: tel-03988667
https://hal.science/tel-03988667v1

Submitted on 14 Feb 2023

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

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

Distributed under a Creative Commons Attribution 4.0 International License

N^o d'ordre : 7172

UNIVERSITE PARIS XI
UFR SCIENTIFIQUE D'ORSAY

THÈSE

Présentée pour obtenir le grade de

DOCTEUR ÈS SCIENCES

DE L'UNIVERSITE PARIS XI ORSAY

Spécialité Informatique

Par

NURIA GALA PAVIA

Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires

Soutenue publiquement le 21 mars 2003
devant la commission d'examen composée de :

M. Christian BOITET (Rapporteur et Président)
M. Eric WEHRLI (Rapporteur)
M. Salah AIT-MOKHTAR (Examinateur)
M. Gregory GREFENSTETTE (Examinateur)
M. Gérard LIGOZAT (Examinateur)
M. Christian JACQUEMIN (Directeur)

Remerciements

J'adresse tout d'abord à Christian Jacquemin mes remerciements les plus chaleureux pour avoir accepté de diriger ma thèse et pour m'avoir accordé un encadrement si dévoué et attentif. Au cours de ces années il a toujours été disponible et ouvert et ses remarques m'ont été precieuses.

Je remercie aussi très sincèrement Gregory Grefenstette qui, depuis mon premier stage à XRCE en 1997, m'a suivie et encouragée dans mon parcours. Merci pour sa confiance témoignée, son enthousiasme et ses conseils toujours enrichissants.

Je tiens à remercier également Salah Aït-Mokhtar pour sa disponibilité, sa rigueur et pour le nombre inestimable de discussions fructueuses que nous avons eues.

Mon travail doit beaucoup à tous les trois et je ne saurai assez leur exprimer ma gratitude.

Je remercie les rapporteurs de ma thèse, Christian Boitet et Eric Wehrli, d'avoir accepté de rapporter mon travail. Je leur suis profondément reconnaissante pour leur lecture minutieuse, critique et constructive de mon manuscrit. La version finale de la thèse leur doit sans doute beaucoup.

Mes remerciements vont également aux membres du groupe CA du Centre de Recherche de Xerox (XRCE). Merci à Claude pour ses nombreux éclaircissements autour de XIP, merci à Hervé, Nicola et Eric pour leurs conseils précieux autour de l'apprentissage. Merci aussi à Caroline B., Caroline H., Agnès, François, Aaron, Bernard et tous les autres pour leurs remarques. Et un grand merci à Victor pour son aide !

Mes remerciements sincères vont aussi à l'ensemble des membres du groupe LIR au LIMSI, pour m'avoir accueillie parmi eux dès le début et pour avoir fait de mes séjours à Orsay des moments riches scientifiquement et humainement. Merci à Laura, Nicolas, Isabelle, Martine, Cécile, Michèle, Anne, Brigitte et tous les autres pour leur amitié.

Je ne peux pas oublier, dans les remerciements, mes parents, ma sœur Silvia ainsi que mes « parrains » Joan et Esperança, qui m'ont toujours motivée, encouragée et aidée le long de mes études. Avec eux, je remercie mon « comité de soutien » : mes grand-mères, mes beaux-parents, mes deux beaux-frères, ma tante Carme et mes amis à Grenoble et Barcelone, pour être à mon côté et partager cette réussite.

Finalement, j'adresse le plus grand des remerciements à mon mari, pour sa patience, sa compréhension et son soutien constant. Sans lui, je n'en serais certainement pas arrivée là !

A Vincent et Magali.

Sommaire

Introduction **11**

I Contexte de l'étude **17**

1 Analyseurs syntaxiques robustes à large couverture **19**
1.1 Introduction . . . 19
1.2 Caractérisation . . . 20
1.2.1 Sortie unique . . . 20
1.2.2 Incrémentalité . . . 21
1.2.3 Empirisme . . . 21
1.2.4 Robustesse . . . 21
1.3 Aspects techniques . . . 22
1.3.1 Approche symbolique . . . 22
1.3.2 Approche statistique . . . 22
1.4 Fondements linguistiques . . . 22
1.4.1 Grammaire Syntagmatique . . . 22
1.4.2 Grammaires de Dépendances . . . 24
1.5 Quelques analyseurs robustes existants . . . 25
1.5.1 Analyseurs suivant une seule approche linguistique . . . 25
1.5.2 Analyseurs mixtes . . . 29
1.5.3 Récapitulatif . . . 32
1.6 Résumé . . . 33

2 Problématique **35**
2.1 Introduction . . . 35
2.2 Le problème de l'hétérogénéité des corpus . . . 35
2.2.1 Textes écrits . . . 36
2.2.2 Transcriptions de l'oral . . . 39
2.3 Évaluation de quelques analyseurs existants . . . 40
2.3.1 Sur des corpus standard . . . 41
2.3.2 Sur des corpus de genres et domaines variés . . . 42
2.3.3 Typologie de phénomènes non (ou mal) modélisés . . . 45
2.4 Le problème des ambiguïtés structurelles . . . 47
2.4.1 Le rattachement prépositionnel . . . 48

2.4.2 Le marquage de la coordination 50
2.4.3 Le repérage de termes complexes 51
2.5 Résumé 53

3 Étude de différents phénomènes 55
3.1 Introduction 55
3.2 Précisions terminologiques 56
3.3 Phénomènes liés à la ponctuation 58
3.3.1 Briève étude de la ponctuation 58
3.3.2 Observations empiriques 61
3.3.3 Parenthèses 62
3.3.4 Guillemets 64
3.4 Listes 65
3.4.1 Les marques 66
3.4.2 L'amorce 67
3.4.3 L'ensemble d'items 70
3.4.4 Relations sémantiques 71
3.4.5 Étude sur corpus 71
3.5 Énumérations 74
3.5.1 Caractéristiques 75
3.5.2 Typologie 76
3.5.3 Observations à partir de l'étude sur corpus 76
3.6 Titres 78
3.6.1 Étude linguistique 78
3.6.2 Étude sur corpus 79
3.7 Résumé 81

4 Vue globale de l'approche 83
4.1 Introduction 83
4.2 Spécifications générales 84
4.2.1 Ajout de pré-traitements spécialisés 86
4.2.2 Définition de deux niveaux d'analyse 88
4.2.3 Lexicalisation des grammaires de dépendances 90
4.3 Présentation de XIP 91
4.3.1 Architecture 92
4.4 Représentation élémentaire des données analysées 93
4.4.1 Arbre syntaxique 93
4.4.2 Système de traits 95
4.4.3 Unités de base 97
4.5 Formalisme d'expression des règles de la grammaire 98
4.5.1 Règles de découpage en syntagmes noyau 99
4.5.2 Règles d'extraction de dépendances syntaxiques 101
4.6 Traitements linguistiques de XIP-F 101
4.6.1 Segmentation 102
4.6.2 Désambiguïsation morphologique 103

4.6.3 Découpage en syntagmes noyau 104
4.6.4 Extraction de dépendances 104
4.7 Résumé 105

II Spécialisation et reconfigurabilité des grammaires 107

5 Pré-traitements et premier niveau d'analyse 109
5.1 Introduction 109
5.2 Pré-traitements 109
5.2.1 Justifications 110
5.2.2 Marquage de listes 111
5.2.3 Marquage de titres 116
5.2.4 Évaluation 116
5.3 La Grammaire Noyau 117
5.3.1 Fondements théoriques 117
5.3.2 Description générale de la grammaire noyau 118
5.3.3 Extraction de dépendances de base 124
5.3.4 Évaluation 125
5.4 Résumé 126

6 Deuxième niveau d'analyse 129
6.1 Introduction 129
6.2 Description générale des grammaires 130
6.2.1 Marquage de phrases N2 130
6.2.2 Correction d'erreurs de chunking 131
6.2.3 Extraction des dépendances 131
6.3 Traitement de la ponctuation 131
6.3.1 Segments spécialisés 132
6.3.2 Dépendances 133
6.3.3 Évaluation du module (ponctuation) 135
6.3.4 Applications possibles 136
6.4 Traitement des titres 138
6.4.1 Segments spécialisés 138
6.4.2 Dépendances 139
6.4.3 Évaluation du module (titres) 141
6.5 Traitement des listes 141
6.5.1 Segments spécialisés 141
6.5.2 Dépendances 145
6.5.3 Évaluation du module (listes) 150
6.6 Traitement des énumérations 153
6.6.1 Segments spécialisés 153
6.6.2 Dépendances 154
6.6.3 Évaluation du module (énumérations) 158
6.7 Récapitulatif 159

6.8 Résumé 160

III Lexicalisation des grammaires de dépendances 163

7 Apprentissage de patrons de cooccurrence 165
7.1 Introduction 165
7.2 Approches existantes 166
7.2.1 Approches linguistiques 166
7.2.2 Approches statistiques 166
7.2.3 Approches hybrides 168
7.3 Aperçu de notre méthode 169
7.3.1 Terminologie 169
7.3.2 Vue générale 170
7.3.3 Annotation des corpus initiaux 173
7.3.4 Critères d'évaluation 173
7.4 Grammaires pour l'extraction de dépendances 173
7.4.1 Grammaires initiales 173
7.4.2 Grammaire de base 174
7.5 Construction d'une base de patrons de cooccurrence 177
7.5.1 Utilisation d'un très grand corpus : le WWW 177
7.5.2 Préparation des requêtes 178
7.5.3 Création d'un grand corpus d'apprentissage 179
7.5.4 Extraction des patrons de cooccurrence 180
7.5.5 Calcul des estimations des probabilités de rattachement 181
7.6 Résumé 182

8 Levée d'ambiguïtés liées au rattachement prépositionnel 183
8.1 Introduction 183
8.2 Levée d'ambiguïtés de rattachement 183
8.2.1 Algorithme 184
8.2.2 Exemple d'application de l'algorithme 185
8.2.3 Évaluation 187
8.2.4 Analyse des erreurs 187
8.3 Variations sur les paramètres 189
8.3.1 Acquisition de dépendances uniquement « fiables » 190
8.3.2 Utilisation de corpus d'apprentissage plus petits 191
8.3.3 Calcul d'un seuil de fréquence 192
8.4 Résumé 193

Conclusion 195

Bibliographie 201

Annexes **213**

A Corpus pour l'étude initiale **213**
A.1 Description 213
A.2 Échantillon des corpus 214

B Traitements linguistiques dans l'analyseur XIP-F **219**
B.1 Analyse morphologique 219
B.2 Analyse syntaxique : découpage en syntagmes 220
B.3 Analyse syntaxique : extraction de dépendances 222

C Scripts de pré-traitements **225**
C.1 Script de balisage de listes 225
C.2 Script de balisage des titres 226

D Grammaires de notre modèle **229**
D.1 Règles de la grammaire noyau 229
D.2 Règles des grammaires de découpage spécialisées 231
D.3 Règles des grammaires de dépendances spécialisées 233
D.4 Spécification des dépendances de base 236
D.5 Spécification des dépendances pour les listes 237
D.6 Règles pour le rattachement prépositionnel 240

E Corpus pour l'apprentissage **243**
E.1 Description des corpus A et B 243
E.2 Échantillon du corpus B 244
E.3 Échantillon analysé 246
E.4 Échantillon corrigé 247

Table des figures

1.1 Représentation par arbre de constituants. 23
1.2 Détail de l'analyse par *chunks*. 24
1.3 Représentation par liens de dépendance. 25
1.4 Détail de la sortie du système FIPS. 26
1.5 Détail de la sortie du système CASS. 26
1.6 Détail de la sortie d'un extracteur de syntagmes nominaux. 27
1.7 Détail de la sortie du système du FDG. 28
1.8 Détail de la sortie du système SEXTANT. 30
1.9 Détail de la sortie du système du GREYC. 31
1.10 Détail de la sortie du système IFSP. 32

2.1 Exemple « simple » de rattachement prépositionnel ambigu. 48
2.2 Exemple « complexe » de rattachements prépositionnels ambigus. 49

3.1 Hiérarchie des marques de ponctuation. 59
3.2 Typologie de structures énumératives. 65
3.3 Éléments composant l'amorce d'une liste. 67
3.4 Amorce avec annexes. 68
3.5 Amorce sans annexes. 68

4.1 Architecture de notre modèle. 85
4.2 Architecture de l'analyseur XIP-F. 93
4.3 Arbre syntaxique de nœuds lexicaux. 94
4.4 Arbre syntaxique de nœuds non lexicaux. 95
4.5 Arbre syntaxique de nœuds non lexicaux avec nœud maximal. 96
4.6 Ensemble de traits associés. 96
4.7 Sortie du segmenteur en unités de base (*phrases*). 102
4.8 Sortie du segmenteur en syntagmes noyau (*chunks*). 104
4.9 Sortie de l'extracteur de dépendances. 105

5.1 Marquage de type HTML. 111
5.2 Marquage avec notre pré-traitement. 112

6.1 Typologie de relations entre des structures syntaxiques. 133

7.1 Architecture générale de notre approche. 171

7.2 Composition de la grammaire G2. 176
7.3 Apprentissage avec le WWW. 178

8.1 Apprentissage exogène avec uniquement des relations `MF1`. 190
8.2 Apprentissage exogène avec des relations `MF1` et `MF2`. 191
8.3 Apprentissage endogène avec uniquement des relations `MF1`. 192

Introduction

Domaine de recherche

Le travail que nous présentons s'inscrit dans le domaine du Traitement Automatique de la Langue (TAL) et plus spécialement dans celui de l'analyse syntaxique robuste (*robust parsing*).

L'analyse syntaxique est un domaine très étudié en linguistique et un composant fondamental dans des techniques informatiques comme la compilation [ASU88]. Les optiques et les objectifs sont très différents selon le domaine : vérifier les prédictions d'une théorie linguistique, tester la grammaticalité d'un code, décrire un langage, etc.

Pourtant, aussi différents que ces objectifs puissent paraître, il existe une réflexion à un niveau général qui est commun : associer à chaque unité en entrée une description structurelle et/ou fonctionnelle, correspondant à une grammaire définie formellement.

Dans une perspective de linguistique computationnelle, depuis des années et spécialement dans les années 70-80, de nombreux formalismes centrés sur la syntaxe sont apparus avec comme objectif une description approfondie des propriétés des structures grammaticales de la langue. Ces formalismes, par exemple TAG [JLT75] et [Jos85], LFG [KB82], GPSG [GKPS85], HSPG [PS87], etc. sont à l'origine d'analyseurs syntaxiques qui vérifient les propriétés décrites dans leurs grammaires tout en donnant à chaque phrase en entrée des traitements syntaxiques approfondis.

Pourtant, l'analyse en profondeur que ces outils fournissent est souvent au détriment de leur robustesse. En effet, on reproche généralement à ces systèmes le fait d'être construits sur la base de « grammaires jouet », capables de traiter des phrases créées artificiellement et dont l'analyse n'a qu'une valeur linguistique.

Ces aspects, liés à d'autres justifications d'ordres théorique et pratique [AMCR02], ont été à l'origine du développement de nouveaux modèles et approches pour l'analyse syntaxique, spécialement à partir des années 90, mais inscrits dans une tradition plus ancienne. En effet, déjà dans les années 70, B. Vauquois [Vau73] avait théorisé le passage « des analyseurs aux transducteurs », seule voie permettant de traiter du texte « réel ». Les travaux de P. Pognan au CERTAL (INaLCO) allaient dans le même sens, ajoutant la contrainte de travailler avec des dictionnaires « minimaux » (voie suivie actuellement par J. Vergne [Ver02]). Dans un cadre plus industriel, on peut citer les travaux chez IBM concernant une grammaire très couvrante de l'anglais (PEG), utilisée de façon opérationnelle pour la correction orthographique, terminologique, grammaticale et stylistique de gros volumes de texte (système EPISTLE, puis CRITIQUE) [Mil80], [MHJ81], et comme

composant du système de traduction automatique SHALT anglais-japonais[1].

Ainsi, du point de vue théorique, il est apparu nécessaire de confronter les hypothèses théoriques avec des données non artificielles (condition *sine qua non* de toute science expérimentale). Du point de vue pratique, l'accessibilité croissante de textes en format électronique a aussi favorisé le développement d'outils capables de traiter des textes réels en grande quantité.

Les outils d'analyse syntaxique de cette approche se placent dans une perspective plus d'ingénierie et de traitement automatique de la langue que de linguistique computationnelle.

Dans cette optique, l'analyse syntaxique consiste à associer automatiquement à la chaîne découpée en unités, une représentation des groupements structurels et des relations fonctionnelles existant entre ces unités, notamment des relations intra et inter syntagmatiques [Abn91, Gre95, AMC97a, KVHA95]. En général, les traitements syntaxiques ne sont pas un but en soi : l'analyse syntaxique est ainsi vouée à la création d'outils permettant l'extraction d'information d'ordre linguistique qui pourra être exploitée par la suite par d'autres applications (par exemple le système FASTUS [AJJ+93] pour l'extraction d'information, GINGER-II [DDTS99] pour la desambiguïsation sémantique ou QALC [FGHP+01], un système de question-réponse).

Ce type d'analyse est aussi fondamental dans des systèmes de traduction automatique (TA). En effet, il correspond au *niveau d'analyse de surface* qui permet d'obtenir une représentation arborescente couvrant tout l'énoncé (structure syntaxique en constituants et/ou graphe de dépendances). La sortie de ce niveau est alors utilisée comme niveau de transfert : une correspondance est établie entre des « sous-arbres » correspondant à des groupes syntaxiques dans la langue source et dans la langue cible [Vau71]. C'est le cas de nombreux systèmes de TA (russe-français et français-anglais au GETA, ALT/JE de NTT, Shalt d'IBM, Mu et Majestic de Kyodai, ATLAS-II de Fujitsu).

Dans la littérature, différentes notions existent pour dénommer l'analyse syntaxique mise en œuvre dans cette approche.

L'*analyse syntaxique de bas niveau* [Deb82, Jen92] est le terme recouvrant l'ensemble des techniques qui extraient un nombre limité de relations syntaxiques d'un texte (*shallow parsing*). Plutôt que produire une analyse linguistique approfondie des phénomènes syntaxiques présents dans un document, ces analyseurs ont comme objectif l'identification des principales structures et/ou relations fonctionnelles, cela de façon plus ou moins détaillée.

À côté de ce terme, la notion d'*analyse syntaxique robuste* [BPL99, AMCR02] (*robust parsing*) met moins en avant la nature de l'analyse linguistique que le fait de produire toujours une analyse pour toute *phrase* en entrée. En effet, les analyseurs dits « robustes » marquent et/ou extraient des structures et des relations fonctionnelles plus ou moins complexes (marquage de syntagmes nominaux, extraction de la dépendance sujet, calcul de dépendances liées au rattachement prépositionnel, etc.), quel que soit le type, le genre ou le domaine du document d'origine (pages Web, journaux, littérature scientifique,

[1] Ces travaux ont été poursuivis en recherche par E. Black à ATR (réalisation d'un très important corpus arboré ATR-Lancaster, faisant suite au corpus IBM-Lancaster, et muni d'informations linguistiques notablement plus fines) et dans l'industrie par le reste de l'équipe d'IBM (G. Heidorn, K. Jensen, etc.) à Microsoft Research après 1992.

encyclopédies, etc.). On peut les appeler, aussi, systèmes *à large couverture*.

Les analyseurs de ce type se veulent donc des outils efficaces pour l'extraction d'informations linguistiques à partir de grands volumes de textes tout venant, sans pour autant être explicatifs des phénomènes linguistiques rencontrés.

Par ailleurs, on constate que l'optique dans laquelle se placent les analyseurs robustes est différente de celle des analyseurs classiques. En effet, ces derniers se trouvent dans une perspective liée à l'intelligence artificielle, leur but étant plutôt la production de toutes les analyses correctes existantes pour un « monde » donné (par exemple celui de la langue naturelle).

En revanche, les analyseurs robustes, se placent dans un paradigme différent, plus lié à la recherche d'information [Gre02]. Dans ce contexte, l'objectif est de produire des analyses pertinentes, en vue d'un besoin particulier (par exemple l'obtention de mots clefs pour indexer un document ou des relations syntaxiques pour interroger une base de données documentaire). L'obtention de ces données linguistiques à partir de grandes quantités de texte « tout venant » passe nécessairement par une analyse linguistique.

Problématique

La notion de « robustesse » (la capacité de toujours fournir une analyse à partir d'un texte tout venant) est fondamentale dans le domaine du traitement automatique de la langue, et plus particulièrement dans l'analyse syntaxique.

Cela présuppose le refus d'une analyse vide même dans le cas de phénomènes non modélisés par les grammaires des analyseurs ou en présence d'entrées mal formées (*ill-formed input*). Dans ces cas, l'analyseur est toujours capable de produire une analyse minimale [AMCR02]. En même temps, étant donné la quantité de texte à traiter, il devient nécessaire de réduire le nombre d'analyses produites (et éviter ainsi la représentation explicite de toutes les possibilités d'analyse d'une entrée ambiguë)[2].

À côté de ces contraintes, un impératif s'impose pour les analyseurs robustes : la production d'une analyse *au moins* partiellement correcte et, surtout, utilisable par la suite dans une application ou tâche automatique.

Il existe des analyseurs robustes majoritairement pour l'anglais[3] (CASS [Abn91], PLNLP [Jen92], ENGCG [KVHA95], etc.), mais aussi des versions pour le français (SEXTANT [Gre95], IFSP [AMC97a], l'analyseur du groupe GREYC à l'université de Caen [GV97b], [Ver02]) ainsi que pour d'autres langues (italien [BPV94], espagnol [GP99], etc.).

L'enjeu auquel ces analyseurs se heurtent est souvent le maintien d'un équilibre entre la finesse de la description linguistique et l'efficacité de l'analyseur [AB99], ainsi qu'entre cette finesse descriptive et leur adéquation empirique.

En effet, la volonté de traiter du corpus tout venant est parfois au détriment d'une analyse globalement précise. C'est-à-dire qu'on retrouve des phénomènes généralement

[2]La possibilité de fournir une analyse vide ou, *a contrario*, de produire un maximum d'analyses pour une même phrase -ambiguë- sont des caractéristiques des analyseurs basés sur les grammaires d'unification, appelés aussi « analyseurs profonds » (*deep parsers*).

[3]Voir http ://www.phil.uni-passau.de/linguistik/linguistik_urls/urls.phtml ?CAT=computing :Software :Parsing

bien modélisés par la plupart des analyseurs (marquage de syntagmes de base –syntagmes nominaux, verbaux, etc.–, extraction de la dépendance sujet et objet, etc.) mais un bon nombre d'autres structures ne sont pas prises en compte par leurs grammaires, ce qui fait diminuer considérablement leurs performances.

Plusieurs raisons justifient cela. D'une part, quelle que soit leur approche, les analyseurs robustes ont souvent été développés en utilisant des corpus standard, principalement des corpus journalistiques, car ce type de corpus a été très tôt accessible électroniquement. Par conséquent, des phénomènes absents ou peu fréquents dans ce type de corpus ont été souvent « oubliés » lors de la création des grammaires [Cha00] (propositions à l'impératif, constructions interrogatives, structures alphanumériques complexes, etc.).

D'autre part, les analyseurs ont parfois négligé des phénomènes traditionnellement peu étudiés en linguistique, à savoir des phénomènes considérées à la frontière entre ce qui est « linguistique » et ce qui a trait à d'autres domaines proches (par exemple la ponctuation [Jon94] ou la disposition visuelle de certaines parties du texte [LGDM+99]).

À ces carences, s'ajoute la difficulté du traitement –avec des moyens uniquement syntaxiques– de structures complexes du point de vue linguistique (notamment le rattachement prépositionnel et la coordination). En effet, ces phénomènes mettent en jeu des ambiguïtés (de rattachement, de portée...) qui peuvent difficilement être bien résolues avec les techniques actuelles.

Objectifs

Tous ces aspects nous ont amenée à réfléchir sur un modèle d'analyseur robuste qui puisse traiter avec précision une grande variété de corpus et de phénomènes, cela tout en conservant une architecture efficace pour le traitement de grands volumes de données. Les aspects linguistiques et informatiques ont été ainsi pris en compte tout au long de notre travail.

Une première partie de notre recherche a consisté à doter l'analyse syntaxique robuste d'un cadre théorique. Pour cela nous avons caractérisé l'existant et nous nous sommes proposé d'établir une description approfondie des phénomènes linguistiques (et textuels) négligés ou mal formalisés par les systèmes actuels.

Dans un deuxième temps, notre objectif a été de proposer (décrire et implémenter) un modèle d'analyseur robuste qui rend compte de la variété des phénomènes présents dans des corpus tout venant (quel que soit leur genre ou leur domaine) et qui garantit la qualité des analyses produites.

Pour ce faire, l'idée de base a été le traitement de l'analyse en deux grandes étapes, selon les caractéristiques de chaque unité en entrée. À la différence de la plupart des analyseurs existants, toute *phrase* est « évaluée » de façon à estimer les types de phénomènes présents, pour être traités par la suite par une grammaire particulière. L'analyseur est ainsi composé de différents modules grammaticaux garantisant un traitement spécifique de chaque phénomène.

De plus, nous nous sommes proposé de modifier les caractéristiques du système initial[4] (système symbolique, écriture manuelle des règles) par l'ajout d'un mécanisme d'appren-

[4] L'analyseur qui a servi de point de départ à notre travail est XIP-F [AMCR02], décrit dans le deuxième partie du chapitre 4.

tissage automatique à partir d'un grand corpus, dans le but de lexicaliser la grammaire de base et d'améliorer ainsi l'extraction de certaines dépendances.

En effet, les systèmes actuels d'analyse syntaxique sont basés sur les parties du discours et n'exploitent que très peu d'informations d'ordre lexical, telles que les rections ou les fréquences des co-occurrences des mots. Cette omission d'informations de type lexical dans une première phase de l'analyse permet d'assurer la robustesse, l'efficacité et la large couverture des analyseurs, mais elle est pénalisante si l'on veut une analyse plus riche et plus précise.

Vue globale de l'approche proposée

Les points précédents ébauchent une vue globale de notre travail. Celui-ci s'articule donc autour de deux notions fondamentales.

a. Spécialisation et réconfigurabilité des grammaires

Les grammaires composant l'analyseur sont appliquées en deux niveaux d'analyse, selon un diagnostic préalable fondé sur les caractéristiques linguistiques et structurelles de chaque phrase en entrée. Il n'y a pas une seule grammaire mais plutôt un ensemble de grammaires (une grammaire noyau et plusieurs grammaires spécialisées pour le traitement de certains phénomènes liés à la ponctuation et à la visualisation de certaines parties du document)[5].

La création de plusieurs grammaires rend l'ensemble modulaire et favorise plus d'efficacité du point de vue informatique. Plutôt que concevoir une seule grammaire dont la taille s'agrandit avec la modélisation de nouveaux phénomènes, nous avons préféré regrouper les règles dans des modules différents, qui s'appliqueront seulement si le phénomène qu'ils décrivent existe dans la phrase en entrée.

b. Lexicalisation des grammaires de dépendances

L'incorporation d'informations d'ordre lexical, puisées automatiquement dans de grandes collections de textes comme le WWW, permet d'enrichir la grammaire et de traiter de façon pondérée l'extraction de dépendances liées au rattachement prépositionnel.

Ces aspects contribuent à la possibilité de produire des sorties munies d'un indice de fiabilité qui pourront être utilisées dans des applications diverses.

Le résultat de notre travail est un modèle d'analyseur robuste et modulaire avec la particularité de donner des indices de fiabilité sur l'ensemble de structures et phénomènes analysés. Devant la difficulté d'obtenir des résultats toujours précis pour des corpus hétérogènes, notre choix a été de donner des informations sur la fiabilité de l'analyse d'une *phrase* traitée par l'analyseur. Ainsi, selon que cette analyse a été calculée par une grammaire ou une autre, le résultat pourra être utilisé différemment dans d'autres applications.

Une telle approche est susceptible d'améliorer l'exploitation des sorties du parseur dans des applications comme la désambiguïsation sémantique ou l'extraction d'informations grâce à la différente pondération des résultats selon leur degré de fiabilité.

[5] Les phrases non *couvertes* par ces grammaires sont quand même analysées, mais la précision de leur analyse est forcément inférieure.

Plan

La première partie présente le contexte de notre étude. Dans le premier chapitre, nous donnons un cadre théorique à l'analyse syntaxique robuste à partir de l'observation de différents analyseurs existants.

Le deuxième chapitre expose la problématique qui a servi de point de départ à notre étude, à savoir, l'hétérogénéité des corpus et les ambiguïtés structurelles. Plus particulièrement, nous évaluons les performances de quelques analyseurs existants devant des corpus hétérogènes et le résultat de cette évaluation nous permet de proposer une première caractérisation de phénomènes linguistiques non pris en compte ou mal modélisés par ces systèmes. Nous examinons aussi trois phénomènes d'ambiguïté, problématiques du point de vue de leur traitement par les analyseurs.

Une étude linguistique et empirique (sur corpus) de quelques phénomènes mentionnés fait l'objet du troisième chapitre. Cette étude nous permet d'une part de traiter en profondeur un certain nombre de phénomènes, et d'autre part de préparer leur modélisation.

Le chapitre quatre présente une vue globale de l'approche que nous proposons et donne des spécifications générales sur la plate-forme qui sert de base à notre travail : le système XIP (*Xerox Incremental Parser*, [AMCR02]). Nous présentons les lignes principales de ce formalisme, à partir de l'analyseur XIP-F existant, et nous mettons l'accent sur les points propres à notre travail.

La deuxième partie de la thèse est centrée sur la notion de spécialisation et de reconfigurabilité des grammaires. Nous présentons dans cette partie deux notions clefs du point de vue linguistique et informatique, respectivement, l'analyse linguistique en deux niveaux et la modularité des grammaires.

Dans le chapitre cinq, nous décrivons les prétraitements originaux que nous avons apportés au système ainsi qu'une grammaire noyau qui est à la base de notre approche. Dans le chapitre six, nous montrons les quelques grammaires spécialisées que nous avons créées au cours de ce travail de thèse. Chaque module est présenté avec des exemples et avec une évaluation détaillée.

La troisième partie, avec les chapitres sept et huit, présente le deuxième aspect clef et original de notre approche, à savoir l'acquisition de ressources lexicales à partir d'une première analyse fournie par l'analyseur pour la levée d'ambiguïtés liées au rattachement prépositionnel. En effet, nous présentons une méthode d'apprentissage de patrons de rection qui permet de pondérer les sorties de l'analyseur concernant le rattachement prépositionnel.

Nous discutons finalement des apports de notre travail et de ses perspectives.

Première partie

Contexte de l'étude

Chapitre 1

Analyseurs syntaxiques robustes à large couverture

1.1 Introduction

La notion d'analyse syntaxique robuste (*robust parsing*) s'est répandue dans les années 90 à la suite d'une effervescence générale pour les techniques informatiques d'analyse de corpus [Cha94]. Cette approche, inscrite dans une tradition plus ancienne ([Vau73], [Mil80], [Deb82]), donne priorité à la robustesse du système, c'est-à-dire à sa capacité de produire une analyse pour tout type de texte en entrée, l'objectif principal étant le traitement de grandes quantités de données.

Les analyseurs de cette approche sont des analyseurs « à large couverture », souvent appelés aussi des « analyseurs de surface » (*shallow parsers* ou *partial parsers*)[1] parce qu'en général ils ne produisent pas une analyse linguistique en profondeur mais plutôt une annotation « minimale » (par exemple marquage de syntagmes nominaux, extraction de la dépendance sujet).

Cependant, on ne peut pas appeler « partiels » (*partial*) ou « surfaciques » (*shallow*) l'ensemble des analyseurs robustes : certains systèmes calculent des relations complexes entre les mots (détection de propositions imbriquées, contrôle, etc.) et la sortie qu'ils produisent ne peut pas être considérée comme une représentation syntaxique « minimale » de la phrase.

Dans ce chapitre, nous présentons une caractérisation générale des analyseurs robustes à large couverture et nous décrivons par la suite quelques analyseurs existants, en mettant l'accent sur leurs fondements linguistiques.

[1] Ces appellations ne sont pas néanmoins synonymes : un « analyseur de surface » ou « surfacique » serait un analyseur qui reconnaît des *chunks* et non pas des constituants classiques ou bien un analyseur ne donnant pas des relations de dépendance. En revanche, un « analyseur partiel » produit une analyse plus riche incluant ce type de relations de dépendance même si certaines sorties sont peu ou mal analysées.

1.2 Caractérisation

De façon générale, un analyseur syntaxique est un outil qui, à partir d'un texte donné en entrée, produit une annotation de ce texte. Selon l'analyseur, les annotations seront plus ciblées sur la structure des éléments intégrant les phrases ou bien sur les relations qui s'établissent entre ces éléments (*cf.* 1.4).

Les analyseurs syntaxiques intègrent des modules de traitement morphologique préalable (segmentation (*tokenizing*), étiquetage (*tagging*) et désambiguïsation morphologique). Les annotations morphologiques produites par ces modules (*cf.* annexe B.1) sont fondamentales pour le calcul de l'analyse syntaxique.

Pour les analyseurs dits « de surface », [Abn94] propose trois caractéristiques principales, à savoir le déterminisme, une évaluation locale des différentes unités de la phrase et l'existence d'une cascade de grammaires spécialisées au traitement d'un type d'unité linguistique (syntagme nominal, syntagme verbal, etc.).

Nous précisons ces idées dans le but de caractériser l'analyse syntaxique « robuste à large couverture ».

- sortie unique : un seul « objet » est produit ;
- incrémentalité : l'analyse de phénomènes linguistiques se fait par étapes ;
- empirisme : les grammaires sont fondées empiriquement (basées sur l'analyse de corpus) ;
- robustesse : traitement de tout type de texte tout venant (une analyse est toujours fournie).

1.2.1 Sortie unique

En général, les analyseurs robustes produisent une sortie unique[2] : un seul « objet » est fourni (objet affiché manipulable du point de vue informatique). Il représente une seule analyse pour la segmentation structurelle (analyse en constituants, par exemple en syntagmes noyau ou *chunks*) mais il peut représenter plusieurs analyses (comme c'est le cas pour le rattachement prépositionnel). Cette analyse se fait par étapes successives et non pas de façon globale et récursive.

L'objectif n'est pas d'explorer toutes les possibles ambiguïtés structurelles ni de fournir toutes les analyses théoriques, mais d'appliquer des heuristiques dans le but de produire l'analyse la plus vraisemblable pour une phrase donnée.

Dans certains cas, seulement quand l'information syntaxique n'est pas suffisante pour appliquer une décision, tous les résultats possibles sont fournis. C'est le cas de l'extraction de dépendances liées au rattachement prépositionnel : le traitement de ce phénomène nécessite des informations plus riches (lexicales, sémantiques) ou obtenues par d'autres techniques (par exemple statistiques) pour établir la bonne relation entre un syntagme prépositionnel et un autre élément dans la phrase dont il dépend (*cf.* chapitres 7 et 8).

[2] Ils sont « déterministes » [Hin83], mais cette dénomination étant très spécifique en théorie d'automates, nous préferons ne pas l'utiliser pour caractériser les analyseurs robustes en général.

1.2.2 Incrémentalité

On peut qualifier l'analyse d'un système robuste de « prudente » : en effet, les décisions difficiles à prendre à un certain stade sont mises en attente. D'après [Eje93], on distingue deux grandes étapes pendant le processus d'analyse :

- une première étape de préanalyse qui a pour objectif la reconnaissance de structures syntaxiques minimales afin de produire une représentation préliminaire qui sera utilisée comme entrée pour l'étape suivante ;
- une deuxième étape dont le but est le calcul de structures plus complexes et/ou des relations entre les mots de la phrase.

Lors de ces deux étapes, l'analyse se fait de façon graduelle, dans des niveaux de granularité différents. C'est pourquoi les analyseurs robustes sont souvent appelés « analyseurs incrémentaux » (*incremental parsers*) : des structures non ambiguës sont analysées dans un premier temps et les résultats de cette analyse initiale sont utilisés pour améliorer les décisions prises dans un stade ultérieur, devant un phénomène plus complexe.

1.2.3 Empirisme

À la différence des analyseurs « classiques » qui reflètent les intuitions linguistiques de leurs auteurs, la formalisation de phénomènes linguistiques dans les analyseurs robustes se fait *a posteriori*. Il s'agit ainsi d'une approche guidée par le corpus plutôt que par une théorie ou formalisme linguistique (*corpus-driven* et non pas *theory-driven*).

La construction de grammaires, faite après l'observation de grandes quantités de corpus, se veut *représentative* de la plus grande partie du contenu syntaxique d'un texte en entrée et non pas *descriptive* de l'ensemble des mécanismes linguistiques existants.

1.2.4 Robustesse

Pour tout type de texte, quel que soit son genre ou domaine ou quelle que soit la structure des items contenus, un analyseur robuste produit toujours une analyse (dont la précision variera considérablement selon les phénomènes modélisés dans la grammaire du système).

En effet, l'analyseur est toujours capable de produire une analyse, même si elle est parfois minimale. Il n'arrive jamais que la sortie soit vide (aucune analyse), même devant des entrées mal formées, erronées ou linguistiquement complexes.

Finalement, il convient de signaler que les résultats obtenus par un analyseur robuste (des corpus annotés syntaxiquement) sont généralement utilisés pour d'autres applications (extraction de terminologie, systèmes de question-réponse, etc.). La réutilisabilité des analyses fournies par un analyseur robuste est aussi une notion importante qui les caractérise.

1.3 Aspects techniques

Par la suite, nous présentons brièvement les principaux paradigmes techniques auxquels appartiennent les analyseurs robustes. D'après [Sri97], on peut distinguer deux grandes approches :

- approche symbolique (*symbolic approach*) ;
- approche statistique (*probabilistic approach*).

1.3.1 Approche symbolique

Les analyseurs symboliques sont constitués de grammaires à base de règles, c'est-à-dire des ensembles de règles qui décrivent des phénomènes syntaxiques. Un nombre important de ce type d'analyseurs est fondé sur la technologie à états finis ([Jos96], [AJJ+93], [Roc93]) mais d'autres utilisent d'autres techniques ([Hin83], [TJ97]).

Dans les analyseurs à états finis, le grammaire est représentée par une cascade de transducteurs créés à partir d'expressions régulières. L'analyse peut être constructiviste si les structures sont ajoutées progressivement au fur et à mesure de l'analyse (cas de [Abn91] et [Gre92]), réductionniste si des restrictions s'appliquent dans le but d'effacer des analyses potentielles qui s'avèrent incorrectes ([CT96], [KVHA95]) ou hybride [AMC97a].

1.3.2 Approche statistique

Les analyseurs de cette approche utilisent des techniques d'apprentissage pour obtenir une représentation syntaxique d'un texte donné. Les règles (ou des poids —probabilités— pour des règles) sont obtenues à partir de textes annotés qui servent à l'entraînement du système (par exemple, pour [Chu88], [DBV99], [MPRZ99]).

Différents modèles statistiques sont proposés dans la littérature (*cf.* 7.2, pour le traitement particulier du rattachement prépositionnel). L'une des difficultés de l'ensemble de ces approches statistiques est le besoin de corpus annotés (manuellement ou automatiquement) pour créer les heuristiques.

1.4 Fondements linguistiques

Selon l'approche linguistique choisie, l'information syntaxique est représentée de façon différente. Il existe deux modèles formels généralisés pour appréhender la réalité linguistique d'une phrase : la première a comme but le marquage de « structures », la deuxième s'intéresse plutôt aux « relations » entre les mots.

1.4.1 Grammaire Syntagmatique

La « Grammaire Syntagmatique » (*Phrase Structure Grammar*) [Blo33] est apparue dans les années 30 aux États-Unis (approche que plus tard a suivi Chomsky pour sa théorie générative transformationnelle).

L'idée fondamentale de cette approche repose sur le principe de constituant immédiat entre les éléments de la phrase, c'est-à-dire que les mots d'une phrase peuvent être regroupés dans des structures plus grandes (en tenant compte de ses significations). L'analyse syntaxique se réduit ici au marquage de catégories et constituants, et à montrer comment ces constituants sont tous structurés au niveau de la phrase.

Le moyen de représentation des informations syntaxiques le plus utilisé dans cette approche sont les arbres.

Voici une phrase en exemple :

(1) *"Une onde ultra sonore est une onde de pression se propageant dans un milieu élastique."*

La figure suivante montre son analyse « classique » en arbre de constituants.

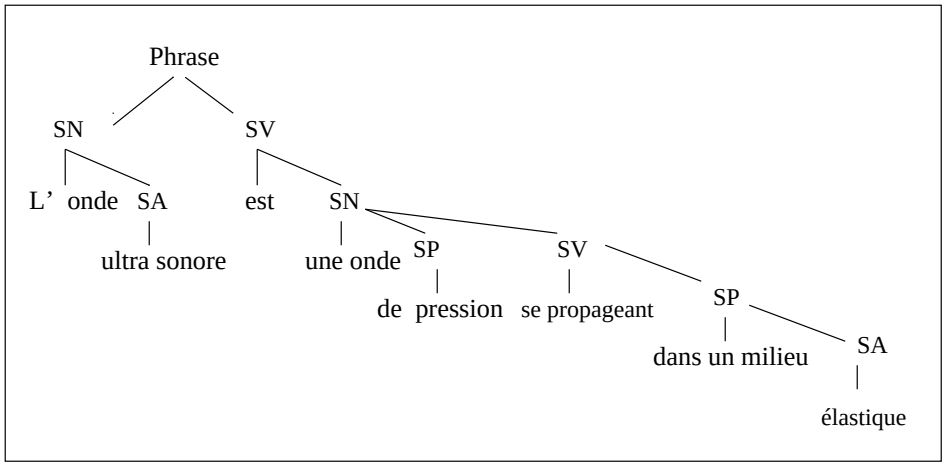

FIG. 1.1 – Représentation par arbre de constituants.

Suivant cette approche fondée sur les structures, Abney [Abn91] propose une décomposition plus fine des structures de la phrase basée sur la prosodie. Le découpage se ferait en « syntagmes noyau » (*chunks*) et non pas en constituants classiques : des ensembles non récursifs des syntagmes classiques motivés par une évidence psycholinguistique.

Dans un *chunk*, l'élément le plus à droite en est la tête syntaxique. [Abn96] en propose une définition précise[3] :

> « Chunk » : le noyau non-récursif d'un constituant dans une proposition, qui s'étend du début du constituant jusqu'à sa tête et qui n'inclut pas des dépendants au-delà de cette tête.

Cet exemple est proposé par le même auteur :

[3] *« A chunk is the non-recursive core of an intra-clausal constituent, extending from the beginning of the constituent to its head, but not including post-head dependents. »*

(2) *The jury said it did find that many of Georgia's registration and election laws are outmoded or inadequate and often ambiguous.*

(« Le jury a dit qu'il a effectivement trouvé que beaucoup des lois d'inscription et d'élection sont démodées ou inadéquates et souvent ambiguës. »)[4]

La figure 1.2 montre la sortie de l'analyse de cet exemple en syntagmes noyau (*chunks*) :

```
[NX The jury] [VX said] [NX it] [VX did
find] that
[NX many] of [NX Georgia]'s [NX
registration and election
laws] [VX are outmoded] or [AX
inadequate] and [AX often
ambiguous] .
```

FIG. 1.2 – Détail de l'analyse par *chunks*.

L'approche de la grammaire syntagmatique a été la tendance prédominante en linguistique jusqu'aux années 70. Parmi les raisons de cette supériorité théorique il y a [Mel88] *« l'utilisation de l'anglais comme langue de base pour l'étude de la linguistique (la structure de cette langue rend le marquage de constituants une tâche relativement facile), l'excessive tendance à la formalisation et une attitude plutôt négligente envers la sémantique »*.

1.4.2 Grammaires de Dépendances

Pendant les années 60, en Europe, une autre approche est apparue avec un intérêt majeur pour la sémantique (par exemple [Fil68] et sa théorie des cas) et pour des langues variées (ayant des structures syntaxiques plus complexes, comme des constituants discontinus). Les « Grammaires de Dépendance » (*Dependency Grammars* [Tes59], [Mel88]) refusent ainsi la représentation stricte imposée par le modèle des constituants. Elles mettent plutôt en évidence les relations établies entre les différents mots de la phrase et comment ces relations sont produites [Kah00].

La représentation de l'information linguistique se fait par des liens entre les mots. Voici l'analyse pour la phrase de l'exemple (1) :

Le grand essor de ce courant théorique se manifeste dans les années 80, ainsi que le montrent les différentes théories apparues à cette période : *Relational Grammar* [Per83], *Word Grammar* [Hud94], *Functional Grammar* [Hal94], *Constraint Grammar* [KVHA95]. L'approche par dépendances a eu une influence sur des théories-formalismes comme LFG [KB82] et HPSG [PS87] où la prise en compte de relations fonctionnelles entre les éléments d'une phrase est une notion fondamentale.

[4]Dans cet exemple, les syntagmes noyaux correspondent à des syntagmes nominaux (`NX`), verbaux conjugués (`VX`) et adjectifs (`AX`). D'autres types de syntagmes existants sont des syntagmes verbaux infinitifs (`INF`), participes présent (`VGX`), adverbes (`RX`).

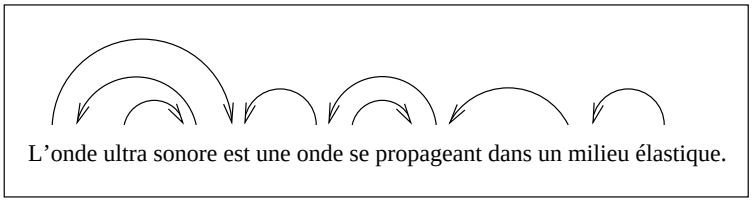

FIG. 1.3 – Représentation par liens de dépendance.

1.5 Quelques analyseurs robustes existants

En tenant compte des approches linguistiques décrites plus haut, il est possible de classifier les analyseurs robustes existants en deux groupes (nous ne caractérisons que les analyseurs symboliques). La tendance actuelle est cependant vers la création d'outils mixtes.

- Le premier groupe est constitué d'analyseurs fondés sur l'une des deux théories : soit la grammaire structurelle, soit la grammaire de dépendances.
- Dans le deuxième, on y trouve des systèmes *mixtes* qui intègrent des aspects des deux approches linguistiques, à savoir, un découpage en constituants et une extraction de dépendances, généralement entre les têtes des syntagmes.

1.5.1 Analyseurs suivant une seule approche linguistique

(A) Approche structurelle

Un système représentatif de cette approche est l'analyseur interactif **IPS** (*Interactive Parsing System*) [Weh92]. Cet outil est basé sur le modèle linguistique de la théorie chomskyenne (simplifiée) du Gouvernement et du Liage (*Government and Binding*). Il s'agit d'un analyseur non déterministe, incrémental ascendant et robuste en ce qu'il arrive à produire des structures partielles même en cas d'échec de l'analyse (constituants minimaux).

Cet outil possède une architecture modulaire avec des modules universaux (communs à toutes les langues) et des modules spécifiques à chaque langue. L'analyse syntaxique prend en compte les têtes lexicales (`N`, `V`, `Adv`...) mais aussi les têtes fonctionnelles (`Tense`, `Infl`...).

La représentation syntaxique se fait par parenthésage (avec la possibilité d'obtenir les étiquettes morphologiques) ou par arbre. La sortie reproduite à la figure 1.4 est obtenue en ligne[5] sur la phrase proposée à l'exemple (1) (la représentation en arbre peut aussi être demandée) :

La couverture linguistique du système FIPS[6] tient compte de structures comme la subordination, la cliticisation, la coordination, les propositions infinitives, interrogatives,

[5] http ://www.latl.unige.ch
[6] FIPS est l'analyseur IPS pour le français.

```
[DP l' [NP onde ]]35
[NP ultra [AP[DP ei ][A sonore ]]]i60
[TP[DP e ][T est [VP [DP une [NP onde [FP[DP e ][F [TP[DP e ]
[T se propageant [VP [AdvP [PP dans [DP un [NP milieu [AP[DP ej ]
[A élastique ]]]j]]]]]]]]]]]]]]402
```

FIG. 1.4 – Détail de la sortie du système FIPS.

etc. Les applications visées par cet outil sont la traduction assistée par ordinateur, l'analyse et l'étiquetage de corpus et l'apprentissage de langues. Il existe des versions pour l'anglais, pour le français, pour l'allemand et pour l'italien.

Le système **CASS** (*Cascaded Analysis of Syntactic Structure*) [Abn90], [Abn91] est composé d'une cascade de transducteurs à états finis qui s'appliquent de façon ordonnée pour reconnaître les constituants de la phrase. Concrètement, l'analyse comprend les étapes suivantes : étiquetage morphologique, marquage de syntagmes, reconnaissance de syntagmes nominaux et identification de certaines fonctions syntaxiques (sujet et prédicat). Chaque transducteur est défini par un ensemble de patrons (catégorie associée à une expression régulière). L'analyseur est déterministe de façon à ne produire qu'une seule sortie par phrase. À notre connaissance, il n'existe qu'une version pour l'anglais.

L'exemple suivant montre la sortie pour la phrase :

(3) *In South Australia beds of boulders were deposited by melting icebergs in a gulf.*

(« En Australie du sud des couches de roches ont été déposées dans un golfe par des icebergs en train de fondre »).

```
Begin-No-Subj
[PP In South Australia]
[Subj [NP beds]]
[PP of boulders]]
[Pred [VP were deposited]]
[PP by melting icebergs]]
[PP in a gulf]]
```

FIG. 1.5 – Détail de la sortie du système CASS.

La sortie est une phrase structurée permettant une meilleure visualisation de l'analyse. Chaque syntagme noyau est entouré de crochets avec l'étiquette correspondant au type de segment et éventuellement la fonction syntaxique (sujet ou prédicat).

Par ailleurs, nous incluons également dans cette approche d'analyse deux types d'outils qu'on ne peut pas considérer comme des analyseurs robustes au même titre que ceux

que nous décrivons dans l'ensemble de la section[7], dans la mesure où ils ne s'intéressent qu'à un certain type de structure (alors qu'un analyseur robuste par segmentation marque tous les constituants présents dans le texte d'entrée). Ce sont des analyseurs de surface très orientés vers des applications particulières.

Le premier type a comme objectif général l'extraction d'information nominale : marquage de groupes nominaux non récursifs (*NP-extractors* ou *NP-tools*) comme celui de [Chu88] et celui de [Sch96] (une sortie de ce dernier est donnée en exemple, figure 1.6). Ils sont utilisés dans des applications comme l'extraction automatique de terminologie.

(4) *Lorsqu'on tourne le commutateur de démarrage sur la position auxiliaire, l'aiguille retourne alors à zéro.*

```
Lorsqu'/CONN on/PRON tourne/VERBP3SG le/DETSG
[ commutateur/NOUNSG de/PREPDE démarrage/NOUNSG ]NP
sur/PREP la/DETSG [ position/NOUNSG auxiliaire/ADJSG ]NP
,/CM l'/DETSG [ aiguille/NOUNSG ]NP retourne/VERBP3SG
alors/ADV à/PREPA [ zéro/NOUNSG ]NP ./SENT
```

FIG. 1.6 – Détail de la sortie d'un extracteur de syntagmes nominaux.

Le deuxième type est spécialisé dans le repérage et l'extraction de certaines structures (en général des noms, des groupes nominaux et parfois des groupes prépositionnels) dans le but d'identifier des entités : termes, noms propres, dates, localisations, etc. Dans la littérature ils sont souvent appelés des « analyseurs légers » (*skimming parsers*) [Jac90]. Nous proposons de les appeler des *analyseurs d'entités nommées*.

Dans ces systèmes les informations extraites sont réorganisées sous forme de patrons d'information avec des variables instanciées. FASTUS [AJJ+93], SCISOR [Jac90], ThingFinder [Tro97], en sont des exemples. Il s'agit d'outils très orientés vers une recherche d'information succincte (souvent en réponse aux questions « qui », « quand », « où »...).

Un outil représentatif de cette approche est **FASTUS** (*Finite State Automaton Text Understanding System*)[8]. Il s'agit d'un système qui vise l'extraction d'informations à partir de texte libre. Il est composé d'une cascade d'automates à états finis et produit plusieurs analyses (non déterministe). Dans un premier temps les expressions syntagmatiques telles que des groupes nominaux, verbaux et prépositionnels sont repérés ; il y a ensuite l'extraction de « patrons » (*event structures*) correspondant à une pseudo-syntaxe (relations sujet et objet).

Les informations syntaxiques permettent d'instancier des variables (par exemple *company : Bridgestone Sports ; activity : production ; year : 2000 ;* etc.). La couverture linguistique est minimale car elle est centrée sur des structures nominales de toutes sortes (incluant des cas de coordination ou d'apposition). L'application de ce système est exclusivement l'extraction d'information, par exemple dans des campagnes d'évaluation

[7] Ils sont des analyseurs robustes mais plus partiels (moins complets) que les autres.
[8] http ://www.ai.sri.com/natural-language/projects/fastus.html

organisées dans le cadre des conférences MUC (*Message Understanding Conference*). Il en existe une version pour l'anglais.

(B) Approche des grammaires de dépendance

À la différence des outils précédents, les analyseurs basés sur les grammaires de dépendance ne segmentent pas les phrases en constituants. Leur objectif est plutôt de rendre compte des différents liens entre les éléments d'une phrase.

Un système représentatif de cette approche est **FDG**, l'analyseur de la grammaire de dépendances fonctionnelle (*Functional Dependency Grammar - English Parser* [TJ97], [Tap99]) de l'université de Helsinki. Il est fondé sur le système ENGCG-2 [VJ96]. Actuellement, cet analyseur est commercialisé pour plusieurs langues par la société Conexor.

Le FDG est constitué de 2700 règles écrites manuellement (en langage C). La description de la syntaxe de surface est basée sur la « Grammaire de contraintes » (*Constraint Grammar*) de [KVHA95]. Le principal élément est un « noyau » (*nucleus*) : le principe de Tesnière est suivi de près. L'élément syntaxique de base est un « nœud » et c'est entre ces types d'éléments que les différentes relations syntaxiques s'établissent. La typologie de dépendances ou « noms de liens » (*link names*) est très variée et s'étend des relations morpho-syntaxiques usuelles (déterminant, sujet, objet, modifieur, etc.) à des relations de nature plus sémantique (temps, but, cause, etc.).

L'objectif visé par cet outil est l'établissement de liens entre tous les éléments de la phrase. La couverture linguistique est sans doute très large ; le FDG est capable de traiter, par exemple, des phénomènes comme l'ellipse verbale, différents types de propositions interrogatives, des propositions-objets, des structures vocatives, des co-prédicats etc.

Il existe une version en ligne pour l'anglais[9] (outils Conexor version 3.4). L'information syntaxique est ici représentée en colonnes.

```
0
1 The           the           det:>2    @DN> DET SG/PL
2 decrease      decrease      subj:>3   @SUBJ N NOM SG
3 reflects      reflect       main:>0   @+FMAINV V PRES SG3
4 the           the           det:>5    @DN> DET SG/PL
5 repeal        repeal        obj:>3    @OBJ N NOM SG
6 of            of            mod:>5    @<NOM-OF PREP
7 catastrophic  catastrophic  attr:>8   @A> A ABS
8 health-care   health-care   attr:>9   @A> N NOM SG
9 benefits      benefit       pcomp:>6  @<P N NOM PL
.
```

FIG. 1.7 – Détail de la sortie du système du FDG.

La première colonne montre le numéro d'occurrence d'un mot dans la phrase, la

[9] http ://www.ling.helsinki.fi/~tapanain/dg/eng/demo.html

deuxième l'item de surface, la troisième son lemme. La quatrième colonne montre les relations de dépendance suivies de la tête syntaxique dont le mot dépend.

Tous les mots sont dépendants d'un autre item dans la phrase. Ainsi le cas du verbe conjugué *reflects* est relié à un élément principal (*main*) qui est le niveau supérieur de la hiérarchie. Finalement, la cinquième colonne affiche l'information morphosyntaxique de chaque unité.

Un autre système représentatif des grammaires de dépendance est l'analyseur **LGP** [ST93] (*Link Grammar Parser*)[10]. La « grammaire de liens » est une théorie développée pour l'anglais où, étant donnée une phrase, des liens étiquetés sont attribués entre des paires de mots. Une nouvelle version de cet outil sortie en 2000 marque aussi les principaux constituants de la phrase.

1.5.2 Analyseurs mixtes

Les analyseurs qui tiennent compte des notions des deux approches linguistiques mentionnées plus haut produisent comme résultat une structure en constituants et extraient des relations de dépendance. La représentation de cette information, ainsi que certaines notions linguistiques, peuvent varier selon les systèmes.

C'est le cas de la notion de « constituant » : il y a des analyseurs qui segmentent le texte en syntagmes traditionnels et d'autres qui se fondent sur la notion de syntagme noyau ou *chunk* [Abn91], [Abn95]. L'analyse selon une approche ou l'autre détermine l'extraction de dépendances.

Nous présentons par la suite quatre systèmes *mixtes*.

Un analyseur de ce type est **CHAOS** (*Chunk Analysis Oriented System*). La première version de ce système est décrite dans [BP92] mais d'autres versions plus sophistiquées ont suivi intégrant, par exemple, des informations sur la rection verbale (*Lexicalized CHAOS*) [BPZ99].

Cette dernière version fait de CHAOS un analyseur non déterministe constitué d'une grammaire à états finis avec des règles lexicalisées. Il utilise un amorçage lexical (*bootstrapping*) dans la mesure où l'acquisition de rections verbales se fait par apprentissage sur corpus.

L'information est représentée par le biais de graphes (les noeuds sont des mots et les branches des dépendances) et les relations extraites sont argumentales (dépendances verbales entre chunks). Les versions de cet outil analysent du texte tout venant pour l'italien et pour l'anglais. Nous ne sommes pas en mesure de reproduire un exemple des sorties de cet analyseur.

SEXTANT [Gre92] [Gre94] est un système déterministe basé sur l'approche des grammaires à états finis. L'analyse syntaxique et le marquage de relations de dépendance

[10] http ://bobo.link.cs.cmu.edu/link/

est focalisé sur les groupes nominaux (et aussi sur d'autres structures comme les groupes prépositionnels ou verbaux contenant des noms où entretenant des relations avec des noms).

Le principal objectif de cet analyseur est l'extraction de patrons syntaxiques par marquage et filtrage comme dans [Deb82]. La notion de *chunk* n'est pas exactement la même que celle d'Abney car les groupes nominaux incluent des groupes prépositionaux (de façon classique).

Il existe sept versions (anglais, français, espagnol, italien, portugais, hollandais et allemand). La représentation de l'information syntaxique se fait avec des crochets pour les constituants de la phrase et une structure parenthésée pour les relations de dépendance.

Exemple pour la phrase :

(5) *Les achats de produits destinés à la revente ne concernent que les négociants, comme la distribution alimentaire ou spécialisée .*

```
PADJ( distribution spécialiser )
NNPREP( produit à revente )
PADJ( produit destiner )
SUBJ( achat concerner )
PADJ( achat destiner )
NNPREP( achat de produit )
-------------------------------------------------
 [NC Les+DET_PL *HeadN achats+NOUN_PL de+PREP_DE *PrepN
produits+NOUN_PL *FreeAdj destinés+ADJ_PL à+PREP_A la+DET_SG
*PrepN revente+NOUN_SG NC] [VC ne+NEG *ActV
concernent+VERB_P3PL VC] que+CONJQUE [NC les+DET_PL
*HeadN négociants+NOUN_PL NC] ,+CM comme+COMME [NC la+DET_SG
*HeadN distribution+NOUN_SG alimentaire+ADJ_SG ou+COORD
*HeadN spécialisée+ADJ_SG NC] .+SENT (1)

Orig:  Les achats de produits destinés à la revente ne concernent
que les négociants, comme la distribution alimentaire ou spécialisée .
```

FIG. 1.8 – Détail de la sortie du système SEXTANT.

Dans [Gre95], la représentation de l'information syntaxique est structurée sous forme de tableau : il y a une colonne pour chaque type d'information produite (numéro de la phrase, type de syntagme auquel le mot appartient, le mot, le lemme, la catégorie, le numéro du mot, etc.) et une ligne par unité lexicale. Les relations de dépendance sont représentées par des chiffres (à la ligne du mot x se trouve l'identificateur numérique du mot avec lequel il entretient une relation). Chaque dépendance a un nom précis (adj, det, subj, etc.).

La couverture linguistique de cet analyseur est relativement réduite car il vise essentiellement les relations entre structures nominales. Les relations concernant des verbes en

forme non personnelle, les propositions interrogatives, les relations entre verbes et adjectifs, etc. ne sont pas traités par ce système dont l'application principale est l'extraction d'informations.

L'**analyseur du GREYC** (Université de Caen) [GV97a] est un système déterministe qui comprend deux niveaux d'analyse, un pour la segmentation en constituants et un autre pour la construction d'une structure fonctionnelle. La représentation de l'information syntaxique est différente selon le niveau mais elle se construit simultanément par un processus d'interaction.

L'analyseur identifie des « groupes non récursifs » (*non recursive phrases*, ou *npr*) correspondant à des syntagmes noyaux et identifie des relations de dépendance entre ces *npr*. Le niveau d'extraction de dépendances utilise des règles et des contraintes de mise en relation (*linking rules* et *linking constraints*) dans une approche basée sur la mémoire : l'analyse est produite par un ensemble de mémoires, une « mémoire » étant une pile d'objets linguistiques.

La figure suivante montre un exemple de la sortie du parseur (obtenue par consultation en ligne[11] sur des phrases et textes choisis par les auteurs) :

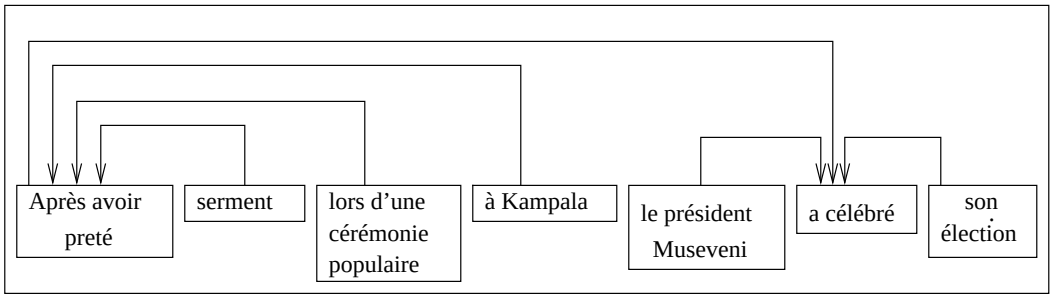

FIG. 1.9 – Détail de la sortie du système du GREYC.

Du point de vue linguistique, ce système couvre partiellement des relations standard comme le sujet et l'objet mais aussi des relations plus complexes comme la coordination, l'énumération et l'ellipse. L'outil de 1997 a été développé pour le français mais les versions suivantes (1998 et 1999) sont multilingues.

Le système **IFSP** (*Incremental Finite State Parser*)[12] [AMC97a] [AMC97b] est composé d'une cascade de transducteurs compilés à partir d'expressions régulières. C'est un analyseur déterministe (au sens d'une seule « sortie » et non pas d'une seule « analyse ») quant au marquage en constituants mais pour l'extraction de dépendances plusieurs relations peuvent être extraites à partir d'un même mot (par exemple lors de l'ambiguïté du rattachement prépositionnel ou des structures coordonnées).

[11] http ://users.info.unicaen.fr/~giguet/syntactic.html
[12] http ://www.xrce.xerox.com/research/mltt/fsnlp/fsparsing.html

Sa stratégie d'analyse générale est constructiviste et réductionniste (des résultats intermédiaires peuvent être supprimés si des conditions attendues ne sont pas vérifiées).

La représentation de l'information syntaxique se fait par parenthésage pour les constituants et pour les dépendances (ces dernières sous la forme `dépendance(x,y)` ou `dépendance(x,préposition,y)` selon le cas). IFSP extrait des relations entre les têtes des syntagmes noyau.

Exemple pour la phrase de l'exemple montré plus haut (5) :

```
SUBJSUBJ(achat,concerner)
DOBJ(concerner,négociant)
PADJ(distributon,alimentaire)
NSURE2([N distribution alimentaire N])
NSURE2([N achat de produit N])
NPPASDOBJ(distributon,spécialiser)
NPPASDOBJ(achat,destiner)
NNPREP(achat,de,produit)
NUNSURE([N  [NP la distribution NP] [AP alimentaire AP] N])
NUNSURE([N  [NP les négociants NP] N])
NUNSURE([N  [PP à la revente PP] N])
NUNSURE([N  [NP Les achats NP]  [PP de produits PP] N])

----------------------------------------------

  Les achats de produits destinés à la revente ne concernent
que les négociants , comme la distribution alimentaire ou spécialisée .

  [SC  [NP Les achats NP]/SUBJ  [PP de produits PP] destinés
[PP à la revente PP] :v ne concernent SC] que [NP les négociants
NP]/OBJ , comme  [NP la distribution NP]/N [AP alimentaire AP] ou spécialisée .
```

FIG. 1.10 – Détail de la sortie du système IFSP.

La couverture linguistique d'IFSP inclut le traitement des propositions enchâssées (*embedded*), le marquage de modifieurs précédant la tête du syntagme, la distinction morphologique de formes verbales (personnelles, non personnelles, passives, réflexives), des relations complexes comme l'apposition et l'extraction de sujets « à longue distance » (partiellement), etc.

L'analyseur a été développé pour le français, pour l'anglais et pour l'espagnol (ce dernier [GP99]).

1.5.3 Récapitulatif

Le tableau suivant synthétise les différents analyseurs décrits jusqu'ici avec leurs caractéristiques principales du point de vue de l'analyse linguistique.

Nom du système	Auteurs	Typologie de l'analyse linguistique
IPS	[Weh92]	structures : constituants
CASS	[Abn91]	structures : syntagmes noyaux
FASTUS	[AJJ+93]	structures : entités nommées
FDG	[TJ97]	dépendances (gram. fonctionnelle)
LGP	[ST93]	dépendances (grammaire de liens)
CHAOS	[BPZ99]	syntagmes noyaux, dépendances
SEXTANT	[Gre92]	syntagmes, dépendances
An. du GREYC	[GV97a]	groupes non récursifs, dépendances
IFSP	[AMC97a]	syntagmes noyaux, dépendances

Table 1. Récapitulatif sur quelques analyseurs existants.

On peut constater qu'une majorité d'analyseurs est fondée sur les grammaires de dépendances, réalisant une analyse structurelle dans une première étape. En effet, cette première étape est souvent d'une grande aide pour le calcul des relations de dépendances (localisation des têtes syntaxiques, délimitation de la portée des constituants, etc.).

1.6 Résumé

Dans ce chapitre, nous nous sommes proposé de caractériser les analyseurs robustes, en mettant l'accent sur leurs particularités par rapport à des systèmes dits « profonds » ou « classiques ».

Notre souci a été, d'une part, de clarifier des aspects concernant la terminologie employée pour faire référence à ces systèmes (« analyseurs robustes », « analyseurs partiels », « analyseurs de surface ») et, d'autre part, de présenter des aspects les caractérisant de façon générale (sortie unique, incrémentalité, empirisme, robustesse).

Nous avons aussi étudié quelques analyseurs existants, spécialement d'un point de vue linguistique.

Une description détaillée de la problématique à laquelle se heurtent ces analyseurs fait l'objet du chapitre suivant.

Chapitre 2

Problématique

2.1 Introduction

Après avoir caractérisé les analyseurs robustes, dans ce chapitre nous allons présenter deux problèmes majeurs auxquels se heurtent ces systèmes quelle que soit leur approche : d'une part, leur comportement face à des corpus hétérogènes et, d'autre part, le traitement des ambiguïtés structurelles.

Dans un premier temps, nous décrivons de façon générale la notion de « corpus hétérogène » tout en donnant des exemples précis rencontrés dans des corpus tout venant. Nous montrons par la suite quelques expériences réalisées avec certains analyseurs existants en utilisant des corpus variés.

Dans un deuxième temps, nous présentons globalement le problème du traitement des ambiguïtés structurelles dans le cadre des analyseurs robustes.

2.2 Le problème de l'hétérogénéité des corpus

Un des problèmes importants pour les analyseurs syntaxiques en général est l'obtention d'une bonne couverture linguistique [AB99] (haut pourcentage de phénomènes traités) sans, d'une part, perdre trop de précision d'analyse et sans, d'autre part, se restreindre à des corpus limités en genre et en domaine (devenant ainsi des systèmes dits *ad hoc*).

D'après les résultats de quelques évaluations réalisées (*cf.* 2.3.1), les taux de précision et de rappel des analyseurs sont satisfaisants sur des corpus du même genre et domaine que ceux utilisés pour le développement de leurs grammaires, notamment des corpus de journaux ou des manuels techniques.

Cependant, le traitement de corpus variés, de genres et domaines nouveaux, implique une baisse dans ces taux (*cf.* 2.3.2). Une des raisons à cela est l'existence d'un bon nombre de phénomènes « non standard » qui ne sont pas pris en compte par les grammaires de ces analyseurs [Cha00].

Plus les quantités de texte tout venant sont grandes et plus les domaines et genres sont variés, plus la présence de phénomènes « non standard » augmente.

Lors de nos premières explorations sur corpus, ayant noté la présence de phénomènes

linguistiques différents selon le mode de production des textes, nous avons séparé transcriptions de corpus oraux et corpus écrits.

2.2.1 Textes écrits

Dans les corpus écrits, nous avons observé plusieurs types de corpus (domaines et genres différents) où nous avons constaté la présence de phénomènes "non standard". En particulier, nous avons identifié les domaines juridique (rapports d'audiences, textes de lois), économique (résultats des marchés de changes, rapports financiers, manuels de type encyclopédique), scientifique (physique, chimie, médecine), technique (modes d'emploi, description de composants), journalistique (*Le Monde*, *Libération*, *L'Humanité*), etc.

À titre d'exemple, voici quelques pourcentages de phénomènes notés sur des corpus variés et qui posent problème aux analyseurs robustes [GP00] (nous avons testé les analyseurs SEXTANT et IFSP, *cf.* 2.3) :

- 5,9 % des erreurs sont dues à des propositions en mode impératif ;
- 4,7 % à des structures avec chiffres et lettres (formules, références bibliographiques, adresses...) ;
- 3,9 % à la présence d'énumérations et de listes ;
- 1,5 % à des propositions exclamatives et interrogatives ;
- 1,1 % à des suites de chiffres ;
- etc.

Un grand nombre de ces problèmes sont liés à la variation typographique. On retrouve plus spécifiquement :

- des suites de chiffres comme des résultats de tennis (*6-4, 6-2*), des numéros de téléphone (*04-76-51-23-44*) ;
- des titres de films/chansons (dans une autre langue) ;
- des adresses ;
- des abréviations ;
- des structures alphanumériques comme des formules (*l'intégrale G1(l,l ')*, *rd=20 mg/cm2*), des références bibliographiques (*Vol. 20, Paris, 1992. ISBN 2-914-0-7*) ;
- des titres et structures sans marque de fin de phrase ;
- etc.

Voici quelques exemples :

(1) *Télécoms : l'italien STET s'allie avec l'américain ATT*

Le groupe italien STET a signé son alliance avec ATT et fait partie désormais du consortium qui unit ATT et Unisource .

Dans ce premier cas, la difficulté principale lors de l'analyse est le manque de ponctuation à la fin de la phrase-titre : les deux phrases sont considérées comme une seule phrase et des liens de dépendance sont extraits de l'ensemble (analyse fournie par le système IFSP [AMC97a]).

```
RELSUBJ(consortium,unir)
SUBJREFLEX(STET,allier)
VMODOBJ(signer,de=le,consortium)
VMODOBJ(signer,avec,ATT)
VMODOBJ(allier,avec,ATT)
ADJ(américain,ATT)
ADJ(italien,STET)
PADJ(ATT,partir)
PADJ(ATT,faire)
PADJ(alliance,partir)
PADJ(alliance,faire)
PADJ(groupe,italien)

[SC  [NP Télécoms NP]/N : [SC  [NP l' italien STET NP]/SUBJ
:v s' allie SC]  [PP avec l' américain ATT PP]  [NP Le
 groupe NP]/OBJ [AP italien AP]  [NP STET NP]/N [v a signé v]
[NP son alliance NP]/OBJ  [PP avec ATT PP] et [AP fait AP]
[AP partie AP] désormais  [PP du consortium PP] [SC [NP qui
NP]/SUBJ :v unit SC]  [NP ATT NP]/N et  [NP Unisource NP]/N .
```

(2) *L'école des hautes études en sciences sociales* **(EHESS, 54 boulevard Raspail, Paris 6)** *accueille du 28 janvier au 6 mars* « **Les images médiatiques et la ville** », *une exposition comprenant 45.000 timbres* **de France, d'Allemagne, d'Espagne et des pays** *de l'ex-bloc de l'Est.*

Dans ce deuxième exemple, le problème réside d'une part dans l'identification de groupes de différentes natures (adresses, titres, énumérations) qu'ils soient entourés ou non de marques de ponctuation (guillemets et parenthèses) et, d'autre part, dans le calcul de liens de dépendance.

```
SUBJ(école,accueillir)
DOBJ(accueillir,ville)
DOBJ(accueillir,image)
VMODOBJ(comprendre,de,est)
VMODOBJ(comprendre,de,bloc)
VMODOBJ(comprendre,de=le,pays)
VMODOBJ(comprendre,de,Espagne)
VMODOBJ(comprendre,de,Allemagne)
VMODOBJ(comprendre,de,France)
VMODOBJ(accueillir,à=le,6_mars)
VMODOBJ(accueillir,de=le,28_janvier)
ADJ(haut,étude)
```

```
PADJ(image,médiatique)

[SC  [NP L' école NP]/SUBJ  [PP des hautes études PP]  [PP en
sciences_sociales PP] ( [NP EHESS NP]/N , 54_boulevard_Raspail ,
[NP Paris NP]/N  [NP 6 NP]/N ) :v accueille SC] [PP du 28_janvier
PP]  [PP au 6_mars PP] <<  [NP Les images NP]/OBJ [AP médiatiques
AP] et  [NP la ville NP]/OBJ >> ,  [NP une exposition NP]/N [v
comprenant v]  [NP 45.000 timbres NP]/OBJ  [PP de France PP] ,
[PP d' Allemagne PP] ,  [PP d' Espagne PP] et  [PP des pays PP]
[PP de l' ex- bloc PP]  [PP de l' Est PP] .
```

(3) *Sur 202 salariés il y a* :

- *70 manœuvres (dont 69 femmes)*
- *44 OS1 (dont 30 femmes)*
- *72 OS2 (hommes)*
- *1 OQ1*
- *16 mécaniciens (hommes)*

Par ailleurs, la pyramide des ages (38-40 ans environ en moyenne) et l'ancienneté (16 ans environ en moyenne) constitue un obstacle au renouvellement et au développement des compétences.

Finalement, dans ce troisième exemple, la difficulté est le traitement de la liste (identification de plusieurs *phrases* à l'intérieur de cette unité, traitement de sa ponctuation ainsi que des espaces horizontaux et verticaux, etc.). En effet, le système IFSP marque six phrases, la dernière comprenant le dernier item de la liste et la phrase suivante.

```
SUBJ(il,avoir)

  [SC  [PP Sur 202 salariés PP] ,  [NP il NP]/SUBJ :v y a SC] : .

   [NP 70 manoeuvres NP]/N ( dont  [NP 69 femmes NP]/N ) .

   [NP 44 OS1 NP]/N ( dont  [NP 30 femmes NP]/N ) .

   [NP 71 OS2 NP]/N (  [NP hommes NP]/N ) .

   [NP 1 OQ1 NP]/N .

SUBJ(ancienneté,constituer)
SUBJ(pyramide,constituer)
SUBJ(mécanicien,constituer)
DOBJ(constituer,obstacle)
VMODOBJ(constituer,de=le,compétence)
```

```
VMODOBJ(constituer,à=le,développement)
VMODOBJ(constituer,à=le,renouvellement)

[SC  [NP 16 mécaniciens NP]/SUBJ (  [NP hommes NP]/N ) Par_ailleurs ,
[NP la pyramide NP]/SUBJ  [PP des ages PP] (  [NP 38-40 ans NP]/N
environ en_moyenne ) et  [NP l' ancienneté NP]/SUBJ (  [NP 16 ans NP]/N
environ en_moyenne ) :v constitue SC]  [NP un obstacle NP]/OBJ  [PP au
renouvellement PP] et  [PP au développement PP]  [PP des compétences PP] .
```

À côté de ces difficultés, les corpus écrits présentent des structures complexes du point de vue de leur traitement linguistique [Mon02] : des propositions interrogatives, exclamatives, à l'impératif, des phénomènes concernant des modifications dans la structure des propositions (vocatives, topicalisation, incises...), des coordinations, des ellipses, d'autres structures comme les corrélatifs, les coprédicatifs (N + N/Adj), les comparatifs, les distributifs, etc.

Tous ces phénomènes, qu'il soient de nature plutôt structurelle ou linguistique, sont distribués différemment selon la typologie de textes [Bib93]. Certains sont propres à un type précis de corpus (abréviations dans les corpus juridiques, symboles et formules dans les scientifico-techniques, propositions interrogatives, exclamatives et impératives dans certains manuels techniques, etc.) alors que d'autres phénomènes apparaissent de façon générale dans les différents domaines (coordinations, mots en langue étrangère, suites de chiffres, titres).

Aussi, l'information textuelle des corpus électroniques est encodée de façon plus ou moins riche : on trouve des textes annotés avec des marques pragmatiques dans les transcriptions orales ou des données appauvries (manque de ponctuation, syntaxe simplifiée par exemple dans les titres, etc.).

Les pourcentages de phrases contenant cet ensemble de structures et paramètres rend nécessaire leur modélisation et leur traitement dans le cadre de l'analyse syntaxique robuste.

2.2.2 Transcriptions de l'oral

Les transcriptions de textes oraux présentent des spécificités telles que nous avons choisi de les présenter séparément.

Du point de vue typographique ou de la structure du corpus, on observe des phénomènes comme le marquage des interventions, le marquage de commentaires de transcription ainsi que des formules conventionnelles pour marquer des blancs, des silences courts, des hésitations, etc.

Pour les phénomènes complexes du point de vue linguistique, nous avons comptabilisé des constructions incomplètes, des ellipses, des phrases contenant des répétitions, des phrases contenant des hésitations, des pauses, etc.

Voici quelques exemples provenant de différents corpus d'*Air France* sur la vente de billets d'avion par téléphone (49.804 mots avec 61 dialogues, 2.161 interventions, 36,2 interventions par dialogue). Le nombre total de mots (toute marque extralinguistique exclue) correspond à 16.692, avec 7,7 mots par intervention, et 22.136 mots en incluant

toutes les marques.

(4) *09 – attendez excusez-moi ... vous êtes intéressé par quel / e par quel vol*
09 – donc e ben je prends 17h quand même

(5) *C21 – <rires> merci beaucoup à vous*

(6) *044 – est-ce que cela ... répond bien à votre question*
C44 – oui oui oui <débit très rapide> ... oui oui disons que ça me conforte un petit peu

La liste suivante montre le pourcentage d'interventions contenant certains des phénomènes fréquents dans les corpus provenant de l'oral :

- 2,5 % commentaires (pragmatiques) ajoutés ;
- 4,7 % hésitations ;
- 6,0 % répétitions.

À cause du manque de ponctuation des transcriptions, nous avons considéré chaque intervention d'un locuteur comme une phrase. Nous avons comptabilisé 2,5 % des interventions présentant des commentaires, 6 % avec répétitions et 26 % de blancs et pauses. La prise en compte de l'ensemble de ces phénomènes n'est donc pas négligeable.

Il semble donc évident qu'un analyseur robuste visant le traitement de ce type de corpus doit présenter des caractéristiques différentes de celles d'un analyseur voué aux textes écrits.

2.3 Évaluation de quelques analyseurs existants

La plupart des analyseurs robustes actuels ont été développés et évalués en utilisant des corpus journalistiques (par exemple, dans l'action GRACE[1] en France [AMP+99], pour l'évaluation d'étiqueteurs morphologiques et analyseurs) ainsi que des manuels techniques. L'ensemble de ces ressources était facilement disponible, notamment avant la généralisation du WWW.

La section suivante présente les résultats rapportés par les auteurs de certains analyseurs, sur les types de corpus mentionnées. Nous avons voulu comparer les performances des analyseurs à notre disposition en utilisant des corpus beaucoup plus hétérogènes.

Les mesures d'évaluation utilisées sont les taux de précision et de rappel classiques. Pour un type d'élément précis, marqué ou extrait par l'analyseur (un syntagme noyau ou une dépendance), ces taux mesurent la proportion d'éléments corrects par rapport au nombre total d'éléments extraits (P) et par rapport au nombre total réel (R)[2] :

$$P = \frac{correctes}{correctes + faux}$$

[1] GRACE : Grammaires et ressources pour les analsyseurs de corpus et leur évaluation, http ://www.limsi.fr/TLP/grace/www/gracdoc.html

[2] La précision est la justesse des résultats trouvés par rapport à la requête originale et le rappel le nombre de résultats trouvés par rapport au nombre de résultats trouvables.

$$R = \frac{correctes}{correctes + manquants}$$

Nous utilisons ces mesures pour évaluer la qualité des analyses fournies par les différents analyseurs étudiés.

2.3.1 Sur des corpus standard

Les évaluations suivantes portent sur des analyseurs pour le français, mais aussi pour d'autres langues à titre de comparaison.

SEXTANT [Gre92]

Parmi les sept versions citées par l'auteur, seule celle pour l'anglais a été évaluée, ceci en utilisant des corpus de manuels de logiciel [Gre95]. Cette évaluation a été menée pour des relations de dépendance binaires (`dépendance(x,y)`) sur 130 phrases du corpus. La table suivante montre les résultats obtenus :

Type de Corpus	*Précision*	*Rappel*
Manuel technique	70 %	64%

Table 1. Taux de précision et de rappel pour SEXTANT (anglais).

D'après l'auteur, les problèmes plus fréquents sont les erreurs d'étiquetage morphologique (*tagging*) propagées le long de l'analyse, certaines marques de ponctuation qui n'ont pas été prises en compte (par exemple des virgules entre noms et verbes). De plus, les propositions interrogatives ainsi que l'extraction de sujets des verbes à l'infinitif n'ont pas été modélisées.

Analyseur du GREYC [GV97b]

L'évaluation de l'analyseur de l'équipe du GREYC pour le français [GV97b] met l'accent sur les relations sujet-verbe, obtenues d'un corpus d'articles du journal *Le Monde* (environ 474 phrases).

Type de Corpus	*Précision*	*Rappel*
Journal *Le Monde*	96,4%	94%

Table 2. Taux de précision et de rappel pour l'analyseur du GREYC (français).

Les erreurs plus importantes sont dues à des constituants (*phrases non récursives*, *cf.* 1.5.2) mal formés, des sujets inversés dans des citations de l'oral, des étiquettes morphologiques mal attribuées ou des coordinations non trouvées.

IFSP [AMC97a]

L'évaluation de l'analyseur IFSP pour le français a aussi été faite pour la relation sujet-verbe sur un ensemble de 157 phrases provenant d'un corpus de domaine technique (manuel) ainsi que sur 249 phrases du journal *Le Monde*.

Type de Corpus	*Précision*	*Rappel*
Manuel technique	99,2%	97,8%
Journal*Le Monde*	92,6%	82,6 %

Table 3. Taux de précision et de rappel pour l'analyseur IFSP (français).

L'analyseur obtient les meilleurs résultats sur des corpus techniques car la grammaire a été développée en utilisant ce type de corpus.

Pour la version de l'anglais en utilisant des corpus journalistiques, le taux de précision pour la relation sujet-verbe est de 84,5 % [AMGP98].

Finalement, la version pour l'espagnol sur la reconnaissance de sujets verbaux sur des corpus techniques (292 phrases) et articles du journal *El Pais* (252 phrases) donne les résultats suivants [GP99] :

Type de Corpus	*Précision*	*Rappel*
Manuel technique	80,7 %	71,8 %
Journal *El Pais*	81,7 %	75,4 %

Table 4. Taux de précision et de rappel pour l'analyseur IFSP (espagnol).

Les erreurs sont principalement des erreurs d'étiquetage morphologique, des coordinations ambiguës ainsi que des groupes nominaux avec des expressions temporelles pris comme de (faux) sujets du fait qu'en espagnol le sujet de surface n'est pas toujours réalisé linguistiquement.

2.3.2 Sur des corpus de genres et domaines variés

Tous les résultats présentés ont tendance à s'affaiblir lorsqu'on traite des textes provenant de domaines différents (par exemple plus spécialisés) : des rapports scientifiques, des transcriptions de procès, des documents économiques, etc.

Cette observation renforce la nécessité de systèmes adaptatifs qui prennent en compte le type de texte [Bib93] ou bien les caractéristiques particulières de certains phénomènes linguistiques. Par exemple [Ill00] montre qu'un étiqueteur entraîné sur des corpus du même domaine obtient de meilleures performances qu'un étiqueteur entraîné sur des corpus journalistiques standard.

Le constat que nous faisons d'après les résultats observés se résume à l'idée suivante :

Même empiriquement fondées (basées sur des corpus), les grammaires des analyseurs décrits jusqu'ici ne sont pas suffisamment représentatives de phénomènes linguistiques et structurels présents dans des corpus spécialisés.

Pour démontrer cette affirmation, nous avons conduit une évaluation sur des corpus hétérogènes en utilisant deux des analyseurs à notre disposition : SEXTANT et IFSP[3]. Il nous a été impossible d'obtenir des résultats du groupe du GREYC concernant leur analyseur.

Les expériences ont porté sur 100 phrases en anglais et 80 en français extraites au hasard de différents corpus spécialisés (voir l'annexe A pour une description plus précise de ces corpus).

L'évaluation concerne le marquage de constituants (notamment la reconnaissance de syntagmes nominaux) et l'extraction de sujets verbaux.

Analyse de constituants

Bien que la plupart des analyseurs disponibles ne présentent pas des résultats d'évaluation sur l'analyse de la structure des phrases (précision du découpage en syntagmes noyaux ou étape de *chunking*), nous avons mené une étude sur les performances de deux analyseurs (SEXTANT et IFSP) pour la reconnaissance de syntagmes, notamment les syntagmes nominaux.

Les résultats obtenus pour l'anglais (E) et le français (F) sont les suivants :

Analyseur	*Précision*	*Rappel*
Sextant E	89,2 %	88,5 %
Sextant F	87,1 %	91 %
IFSP E	87,7 %	83,3 %
IFSP F	90 %	91,4 %

Table 5. Taux de précision et de rappel
pour la reconnaissance de syntagmes nominaux.

Une remarque s'impose tout d'abord : la définition de « syntagme » est différente pour les deux analyseurs. Dans SEXTANT, le syntagme correspond glabalement à la notion classique de « groupe », qui inclut la tête du syntagme ainsi que ses modifieurs.

Voici le découpage de SEXTANT (la sortie montre aussi les étiquettes morphologiques) pour la structure *« Les achats de matières premières »* dans la phrase :

(7) *Les achats de matières premières concernent surtout les industries de transformation.*

```
[NC Les+DET\_PL *HeadN achats+NOUN\_PL de+PREP\_DE *PrepN matières+NOUN\_PL
*FreeAdj premières+ADJ2\_PL NC]
```

[3][Mon02] a fait un travail similaire en comparant les analyseurs CASS [Abn90], IPS [Weh92], IFSP [AMC97a] et l'analyseur de la grammaire de liens (*Link Grammar*) de [ST93] ; les résultats sur des corpus particuliers montrent quelques phénomènes linguistiques qui posent problème à ces analyseurs.

En revanche, dans IFSP la notion de syntagme est plutôt celle de « syntagme noyau » (*chunk*) où l'élément le plus à droite est la tête du syntagme. Le découpage suivant est celui d'IFSP pour le même exemple :

```
[NP Les achats NP] [PP de matières PP] [AP premières AP]
```

Les erreurs de précision dans les deux analyseurs proviennent souvent d'un traitement morphologique imprécis (que ce soit à la segmentation —*tokenisation*— ou à l'étiquetage morphologique —*tagging*). Une autre source importante d'erreurs est la ponctuation, notamment, la reconnaissance de constituants inappropriés dans des cas de séries de nombres ou listes de noms avec des virgules ou points et virgules.

En ce qui concerne le rappel, certains des syntagmes nominaux non identifiés correspondent à des mots mal étiquetés (ambiguïtés morphologiques) ainsi qu'à des titres ou items de listes sans ponctuation (et marqués comme s'ils faisaient partie d'un unique constituant).

Extraction de la dépendance sujet

La table suivante montre les résultats de l'évaluation pour l'extraction de sujets :

Analyseur	*Précision*	*Rappel*
Sextant E	70,8 %	63,8 %
Sextant F	65,1 %	44,3 %
IFSP E	78,9 %	74,8 %
IFSP F	82,3 %	86,8 %

Table 6. Taux de précision et de rappel pour l'extraction de la dépendance sujet.

Pour l'anglais, les résultats de SEXTANT sont légèrement meilleurs que ceux rapportés par l'auteur (l'évaluation originale avait été menée sur des corpus de domaine technique contenant des phénomènes "complexes" du point de vue de l'analyse syntaxique automatique : listes, propositions impératives, etc.).

Les résultats d'IFSP sont moins bons pour la précision que ceux obtenus sur les corpus journalistiques, notamment à cause de la ponctuation (guillemets) et de quelques structures ambiguës (coordinations).

Pour le français, il n'existe malheureusement pas d'évaluation publiée pour SEXTANT. Les erreurs que nous avons remarquées proviennent essentiellement de la segmentation, de l'étiquetage morphologique, ainsi que de la non-identification de sujets inversés.

Les résultats obtenus avec IFSP pour les deux langues sont moins précis à cause des même problèmes (segmentation, étiquetage morphologique) ainsi qu'à de fausses relations établies entre items de phrases différentes (qui n'avaient pas de ponctuation finale, par exemple les titres) et que le parseur n'a pas identifiées. Le rappel est aussi moins bon comparé à l'évaluation des corpus de journaux et des corpus techniques (*cf.* table 3).

Pour conclure cette partie, la table 7 donne un récapitulatif des deux analyseurs pour l'anglais et le français en utilisant des corpus différents. La première colonne montre l'évaluation citée par les auteurs (pour IFSP français il s'agit d'une moyenne des précisions de la table 3). La deuxième colonne présente les résultats que nous avons obtenus dans des corpus spécialisés contenant une plus grande variété de phénomènes :

Analyseurs	*Corp. journ. / techn.*	*Corp. Spécialisés*
Sextant E	**70 %**	**70,8 %**
Sextant F	(-)	**65,1 %**
IFSP E	**84,6 %**	**78,9 %**
IFSP F	**95,2 %**	**82,3 %**

Table 7. Moyennes des taux de précision de deux analyseurs sur des corpus variés.

Pour l'analyseur SEXTANT, la précision augmente légèrement (il faut tenir compte du fait que des corpus techniques avaient été utilisés lors de la conception de sa grammaire).

Pour IFSP, les résultats sont clairs : pour les deux langues, il y a une baisse significative de la précision lorsque des corpus spécialisés sont pris en entrée.

2.3.3 Typologie de phénomènes non (ou mal) modélisés

D'après les observations réalisées, nous avons voulu faire une première typologie des phénomènes présents dans des corpus spécialisés qui posent problème aux analyseurs, soit parce qu'ils sont mal modélisés, soit parce qu'ils ne sont simplement pas modélisés du tout par les grammaires.

Dans des corpus hétérogènes, ce sont principalement des phénomènes que nous considérons liés à des marques non lexicales (ponctuation, marques visuelles, etc.) comme des propositions interrogatives et exclamatives, des « unités délimitées » (entourées de marques de ponctuation spécifiques), des listes (énumérations disposées verticalement) et des unités liées à la structure du document (titres de différents types). Il y a aussi des phénomènes imposant des contraintes sur l'organisation générale de la structure de la phrase (absence de sujet de surface dans les propositions impératives).

Propositions en mode impératif

La caractéristique principale des propositions en mode impératif est le manque de syntagme nominal sujet. La création d'une règle dans le parseur qui modélise cette connaissance évitera l'extraction de tout syntagme nominal postérieur au verbe comme sujet inversé (plutôt qu'objet).

(8) **Refermez** *l'ouverture latérale.*

(9) **Abaisser** *le bouton de commande pour l'engagement.*

Propositions interrogatives et exclamatives

Ces propositions présentent des marques de ponctuation finales spécifiques (« ? » ou « ! ») et peuvent être constituées d'un ou plusieurs syntagmes ou bien elles peuvent être des phrases complètes. Dans le cas des propositions interrogatives, l'ordre des éléments varie : le verbe précède le syntagme nominal sujet.

Exemples[4] :

(10) *Halte au massacre du peuple palestinien* **!**

(11) *Quel est son véritable programme* **?**

Les relations syntaxiques à extraire doivent prendre en compte l'absence de sujet dans certains cas, ou l'ordre spécifique des constituants.

Unités délimitées

Nous considérons comme « unités délimitées » des entités entourées de marques de clôture (*cf.* 3.3.1), principalement parenthèses, crochets, tirets, guillemets français (« »), guillemets anglais (“ ”) ou apostrophes (' '). Selon leurs caractéristiques, elles sont généralement appelées citations, entités nommées (titres de livres, films, etc.), commentaires, appositions, dialogue transcrit, etc.

Ces unités apparaissent séparées du texte principal dans la mesure où elles sont entourées des marques de ponctuation ouvrantes et fermantes. Du point de vue de leur structure morpho-syntaxique, il peut s'agir de sigles, acronymes, mots, syntagmes, groupes de syntagmes ou phrases entières. Exemples :

(12) *Parmi les autres valeurs à la baisse, on remarquait Pinault-Printemps, dont les ambitions dans le secteur* —**très gourmand en capitaux**— *des télécommunications semblent inquiéter le marché.*

(13) *Le thème* « **Structure et Dynamique des Atomes et des Ions** » *a une place toute particulière dans les activités du laboratoire.*

(14) *Il évoque aussi, dans l' Express, les* **'disparitions'** *et ce qu'était l'état d'esprit des chefs militaires de l'époque.*

L'annotation syntaxique de ces segments doit prendre en compte ces structures en tant que constituants propres, indépendamment de leur structure interne. Une telle annotation permettra une meilleure extraction des dépendances par la suite.

Listes

Comme nous allons le montrer en détail plus loin (*cf.* 3.4), les listes sont des énumérations d'items organisés verticalement (*“vertical lists”* [Whi95]). Elles sont composées d'une séquence introductrice (amorce) ainsi que d'une série d'items introduits par des marques de ponctuation spécifiques (séparateurs ou organisateurs) :

[4] Tous les exemples sont extraits des corpus décrits à l'annexe A.

(15) *3 - Desserrez la vis - pointeau de purge (un 1/2 tour à un tour) .*
4 - Appuyez sur la pédale de frein
5 - Fermez la vis - pointeau de purge (...).

(16) *Le présent chapitre analyse certaines des possibilités commerciales existant au chapitre :*
- des céréales ;
- des oléagineux ;
- des légumineuses à graines ;
- [...]

Le traitement syntaxique de ces unités implique une bonne reconnaissance initiale de l'ensemble de la liste (*cf.* 5.2).

Unités liées à la structure du document

Il s'agit en général de titres de sections ou sous-sections (*cf.* 3.6) apparaissant sans des marques de ponctuation « classiques » mais séparées du reste du texte par des marques visuelles (verticales et parfois horizontales). Ces unités sont constituées de noms, de groupes nominaux (principalement) mais peuvent être aussi des propositions complètes :

(17) *TABLE DES MATIÈRES :*

(18) *1. APERÇU DE L'AFRIQUE DU SUD*

(19) *L'emploi américain fait sortir le dollar de l'ornière*

La principale difficulté pour l'analyse syntaxique est d'effectuer un repérage initial correct de ces unités (*cf.* 5.2).

Une étude plus détaillée de certains des phénomènes présentés dans cette section est donnée au chapitre 3.

2.4 Le problème des ambiguïtés structurelles

De façon générale, une des difficultés majeures de tout système d'analyse syntaxique automatique est le traitement des ambiguïtés au niveau de la phrase. Les ambiguïtés structurelles les plus importantes concernent des phénomènes comme le rattachement prépositionnel, la coordination et le repérage de termes complexes. Le traitement de ces phénomènes implique la prise en compte des ambiguïtés de rattachements des constituants.

2.4.1 Le rattachement prépositionnel

Le rattachement prépositionnel est le « cas canonique » de l'ambiguïté structurelle [HR93]. Le problème concerne le lien de dépendance entre un syntagme prépositionnel (`SP`) et une tête syntaxique qui peut être un verbe, un nom ou un adjectif. Selon la position du `SP` (notamment quand il se trouve après un verbe) il n'est pas toujours possible de déterminer quel est le bon rattachement sur la base de la structure morpho-syntaxique :

(20) *Les variations soumettent les particules* **à des mouvements** ...

(21) *Les variations soumettent les particules* **de faible résistance** ...

En effet, les deux phrases précédentes ont la même structure syntagmatique (`SV SN SP`) ; mais dans (20) le `SP` se rattache au verbe (*soumettre à mouvements*) alors que dans (21) le `SP` se rattache au nom (*particules de faible résistance*).

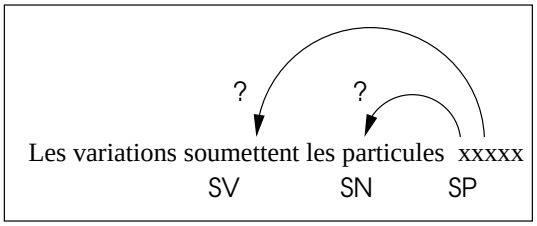

FIG. 2.1 – Exemple « simple » de rattachement prépositionnel ambigu.

De plus, les possibilités de rattachement augmentent avec le nombre de constituants dans la phrase :

(22) *Il s'agit [de la propagation] [d'une énergie] mécanique [dans un milieu] matériel.*

Pour cet exemple, qui présente la configuration `SV SP SP SA SP SA`, le premier `SP` de la phrase se rattache au verbe sans ambiguïté. En revanche, *dans l'absolu et d'un point de vue syntaxique*, le deuxième `SP` « d'une énergie » peut être rattaché au verbe « agir » mais aussi au nom du `SP` précédent (« propagation »). De même, le `SP` « dans un milieu » peut être rattaché au verbe, au nom du premier ou du deuxième `SP` et aussi à l'adjectif « mécanique » (voir figure 2.2). Six rattachements sont donc possibles dans pour ces deux `SPs`, alors qu'en l'occurrence seules les relations « propagation d'une énergie » et « propagation dans milieu » seraient ici correctes.

Il est certain que l'ajout d'information linguistique plus riche (par exemple, la distinction entre arguments et circonstants –ajouts– ou l'utilisation de traits sémantiques) améliorerait la résolution des rattachements. Dans l'exemple, le lien « agit d'une énergie » serait d'office éliminé car le verbe « agir » régit un argument de_N, en l'occurrence « d'une propagation ». Ce `SP` ne peut pas être un circonstant car la préposition « de » introduit un circonstant de matière, ou de localisation temporelle ou spatiale alors que les traits sémantiques de « propagation » ne le permettent pas.

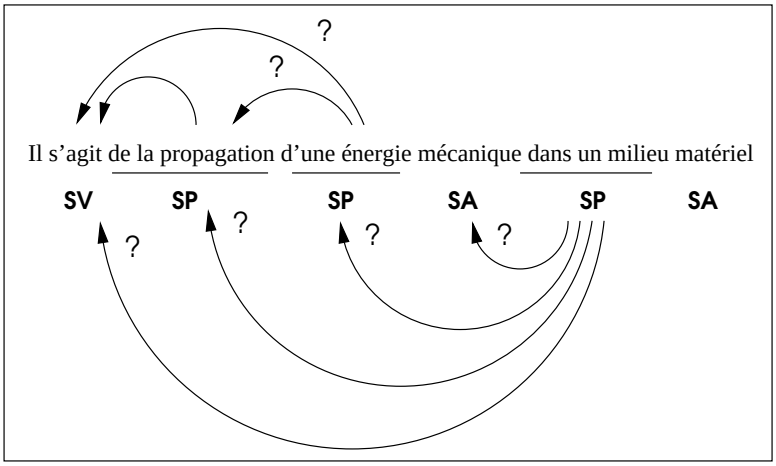

FIG. 2.2 – Exemple « complexe » de rattachements prépositionnels ambigus.

Cependant, l'accès à de telles informations linguistiques n'est pas toujours possible (il n'existe pas à l'heure actuelle pour le français une base lexicale complète et facilement accessible contenant des informations affinées sur verbes, noms et adjectifs [BF00]). Dans ce cas, le système doit être capable de trouver le bon rattachement avec les informations dont il dispose.

Pour ce faire, un bon nombre d'analyseurs syntaxiques existants effectuent tout d'abord un découpage en syntagmes (*cf.* 1.5.2) et, dans un deuxième temps, cherchent à résoudre les ambiguïtés de rattachement en utilisant les têtes de ces syntagmes. Or, la tâche de découpage ne s'avère pas toujours optimale dès qu'il y a des groupes prépositionnels en jeu. Ainsi, d'après les résultats obtenus par [Mon02], l'identification de groupes prépositionnels par rapport à l'identification de groupes verbaux est globalement moins bonne : chez SCOL-CASS (anglais) [Abn96] 74,2 % de précision pour les `SP` et 85,6 % pour les `SV` ; chez IPS 80 % pour les `SP` et 98,5 % pour les `SV`.

En ce qui concerne l'extraction de dépendances, une évaluation de l'analyseur IFSP (anglais) [AMGP98] montre que des relations de dépendance comme le sujet obtiennent de meilleurs résultats que des dépendances liées au rattachement prépositionnel (il faut prendre en compte le fait que la stratégie appliquée pour l'extraction de relations de dépendance liées au rattachement prépositionnel est non-déterministe). Les rattachements dans IFSP [AMC97a] sont représentés par les relations `NNPREP` pour le rattachement au nom et `VMODOBJ` pour le rattachement au verbe. La relation sujet (`SUBJ`) a un taux de précision de 84,53 % alors que `NNPREP` seulement 53,37 % et `VMODOBJ` 65 %.

Voici quelques exemples avec des relations erronées extraites par IFSP :

(23) *Several suspected thieves are juveniles hired by organized rings across the border, Flores said.*

(« Plusieurs voleurs suspects sont des jeunes payés par des bandes organisées de l'autre côté de la frontière, a dit Flores ».)

```
NNPREP (ring,across,border)
NNPREP (juvenile,by,ring)
```

(24) *Joe Padilla, an Army spokesman, said he had no information on the circumstances of the accident.*

(« Joe Padilla, un porte-parole de l'armée, a dit qu'il n'avait pas d'informations sur les circonstances de l'accident ».)

```
VMODOBJ(have,on,circumstance)
```

(25) *Facing a possible challenge from Jesse Jackson, Washington Mayor M. Barry is heading (...)*

(« Confronté à un possible défi venant de Jesse Jackson, le maire de Washington, M. Barry, mène (...) »)

```
VMODOBJ(face,from,Jackson)
```

Différentes méthodes sont proposées dans la littérature pour résoudre les ambiguïtés de rattachement dans le cadre de l'analyse syntaxique automatique. Nous en donnons une description détaillée, avec une proposition personnelle, dans la troisième partie de ce rapport.

2.4.2 Le marquage de la coordination

Un des problèmes d'ambiguïté structurelle les plus fréquents (avec le rattachement prépositionnel) est le marquage des coordinations. Une coordination unit deux éléments (unités lexicales, syntagmes, groupes de syntagmes ou propositions entières) de façon à ce que les éléments coordonnés réalisent la même fonction syntaxique.

La nature des éléments coordonnés peut être identique (par exemple deux syntagmes nominaux) ou bien différente (adjectif et participe passé, syntagme nominal et proposition subordonnée, etc.).

Les ambiguïtés au sein des coordinations peuvent porter sur l'élément coordonné lui-même (26) ou sur le rattachement d'un élément extérieur à la coordination (27) [Bru98] :

(26) *[[Une pomme et une poire] ou une fraise]*

[Une pomme et [une poire ou une fraise]]

(27) *Il achète [[des livres et des cahiers] neufs]*

Il achète [des livres et [des cahiers neufs]]

De plus, il existe des phénomènes très variés liés à la coordination.

L'**ellipse** est une coordination où l'un des éléments l'intégrant, souvent le deuxième, est absent (le sujet dans l'exemple suivant) :

(28) *Le débat concernant la nature juridique de la cession de 1940 [[est théorique en l'espèce] et [n'a pas à être tranché]].*

Dans les **distributions** il y a deux coordinations en parallèle ; les éléments sont liés le premier avec le premier et le deuxième avec le deuxième (dans l'exemple suivant, *« importés »* avec *« à 4,3 »* et *« exportés »* avec *« à 5,45 milliards de rands »*). On retrouve dans la plupart des cas un adverbe ou un adjectif comme *« respectivement »*, *« respectifs »*, etc.

(29) *La valeur totale des produits agricoles [[importés] et [exportés]] a été estimée, en 1993, [[à 4,3] et [à 5,45 milliards de rands], respectivement].*

Lors des **corrélations**, deux éléments coordonnés apparaissent avec un élément de coordination spécifique qui les précède : *ni... ni...*, *soit... soit...*, etc.

(30) *La Loi sur l'éducation n'exige pas que les enfants fréquentent [[soit une école laïque] [soit une école catholique romaine]].*

Les **énumérations**[5] sont des coordinations contenant plus de deux éléments séparés par des virgules, sauf l'avant-dernier et le dernier qui sont généralement séparés par une conjonction de coordination. Parfois, elles présentent une marque de ponctuation spécifique (deux points précédant les éléments énumérés, des parenthèses) :

(31) *Présents : [les juges [La Forest], [L'Heureux-Dubé], [Sopinka], [Gonthier], [Cory], [McLachlin], [Iacobucci] et [Major]].*

(32) *Le réglage s'effectue dans l'ordre numérique des cylindres : [1, 2, 3 et 4].*

(33) *Après avoir accédé aux salles d' information, les cinq candidats [([Groupama], [Swiss Life], [Eureko] et [les deux américains [AIG et GE Capital]])] qui ont manifesté leurs intérêts pour l'assureur public sont à l'heure des choix.*

L'identification automatique de la portée des éléments coordonnés exige un traitement spécifique au niveau du marquage en constituants et de l'extraction de dépendances. Quelques indices peuvent aider à identifier des éléments coordonnés : symétrie de la construction, absence ou présence homogène de déterminants, prépositions identiques, ponctuation, certaines restrictions sur les arguments sélectionnés... [Bru98].

2.4.3 Le repérage de termes complexes

Les termes sont « des objets linguistiques utilisés dans la littérature technique et scientifique et visant à faire référence à des concepts de façon non ambiguë et consensuelle » [Jac97].

On distingue traditionnellement les termes simples des termes complexes. Les premiers sont des unités lexicales pleines (des noms) alors que les deuxièmes sont des syntagmes

[5] Étant donnée leur similitude avec les listes, nous allons étudier les énumérations plus en détail (*cf.* 3.5).

nominaux composés d'au moins deux unités lexicales pleines qui, de par leur structure, demandent une analyse syntaxique fine [Mor99].

Les termes complexes sont souvent assimilés à des noms composés[6]. On y retrouve une tête avec des modifieurs, organisés selon des contraintes d'ordonnancement assez rigides [Jac97]. Ils sont composés d'au moins un nom et peuvent contenir un adverbe, une préposition, plusieurs adjectifs et éventuellement une coordination : *monnaie unique* (`N Adj`), *commutateur d'éclairage* (`N Prép N`), *berges de résection microscopiquement envahies* (`N Prép N Adv Adj`), etc.

Plus le terme est complexe, plus on retrouve des problèmes de découpage terminologique. Dans [Jac97], l'auteur donne l'exemple de deux termes ayant une même structure morphologique (`Adj N N`) mais présentant une structure syntagmatique différente :

[natural language] [processing] (« traitement (automatique) du langage naturel »)

[dynamic] [information processing] (« traitement dynamique de l'information »)

Le parenthésage rend évident la différence qui se manifeste aussi lors de la traduction.

Les ambiguïtés structurelles concernent principalement la portée de la modification adjectivale ou nominale, parfois agravée par le phénomène de la coordination [Man01]. Les deux exemples suivants présentent aussi la même structure (deux noms coordonnés, le premier avec un modifieur `Adj` et le deuxième avec un modifieur `N`) mais le terme est différent dans les deux configurations :

[[angiographic equipment] and techniques] (« appareils de coronarographie et techniques »)
[catheter [occlusion or infection]] (« occlusion ou infection de cathéter »)

La résolution de ce type d'ambiguïté nécessite un apport important d'information lexicale dépendant du sous-langage de spécialité et difficilement généralisable à d'autres domaines.

Il existe différents outils dédiés à l'extraction de termes et de leurs variantes. Comme pour les analyseurs robustes (*cf.* 1.3) on distingue des méthodes symboliques (FASTR [JR94], LEXTER [Bou94]) et des méthodes statistiques [Chu88]. Il existe aussi des méthodes qui combinent les deux approches, symbolique et numérique [Dai94].

Comme c'est le cas pour la désambiguïsation du rattachement prépositionnel, des moyens uniquement syntaxiques s'avèrent insuffisants pour le repérage des termes. Des analyseurs comme IFSP adoptent une stratégie non déterministe dans le but d'extraire un maximum d'occurrences (parmi lesquelles il y a des candidats termes). Par exemple :

(34) *Elle représente un marché en croissance à la fois vaste et très prometteur en particulier pour ce qui est des aliments à caractère ethnique .*

```
NUNSURE( pour_ce_qui_est_des aliments à caractère ethnique )
NUNSURE( un marché en croissance à_la_fois vaste et très prometteur )
NUNSURE( un marché en croissance à_la_fois vaste et très prometteur
    en_particulier pour_ce_qui_est_des aliments à caractère ethnique )
```

[6]Les termes se distinguent des noms composés principalement par l'absence de déterminant.

(35) *Ces grandeurs sont des données indispensables tant en astrophysique qu'en physique des plasmas ou même en physique nucléaire .*

```
NUNSURE( en physique_nucléaire )
NUNSURE( en physique des plasmas ou même en physique_nucléaire )
NUNSURE( en astrophysique qu'en physique des plasmas ou même en
    physique_nucléaire )
NUNSURE( des données indispensables tant en astrophysique
    qu'en physique des plasmas ou même en physique_nucléaire )
```

Dans IFSP, la relation `NUNSURE` extrait des groupes de syntagmes candidats à unités terminologiques. Le taux de précision de cette relation n'a pas été évalué mais, à l'évidence, les heuristiques mises en œuvre ne s'avèrent pas suffisantes.

2.5 Résumé

Dans ce chapitre, nous avons présenté une étude des principaux problèmes touchant à l'analyse syntaxique robuste.

D'une part, nous avons montré que l'objectif ambitieux de traiter des corpus tout venant de domaines et de genres hétérogènes implique une baisse des performances des analyseurs robustes (en termes de taux de précision) en ce qui concerne l'analyse linguistique fournie. Les résultats de l'évaluation de deux analyseurs robustes existants démontrent empiriquement cette affirmation qui se justifie par la difficulté (ou impossibilité) de modéliser dans une seule grammaire l'ensemble des phénomènes présents dans ces corpus.

Après avoir exploré les raisons de cette baisse des performances, nous avons aussi décrit brièvement des phénomènes liés à des marques non lexicales (ponctuation, marques visuelles, etc.) qui de façon générale ne sont pas pris en compte par les grammaires des analyseurs existants ou bien qui sont mal modélisés.

D'autre part, nous avons présenté globalement la problématique des ambiguïtés structurelles que nous avons illustré, avec le cas du rattachement prépositionnel, de la coordination et du repérage de termes.

Dans le chapitre suivant, nous présentons en détail quelques phénomènes liés à la ponctuation et à la structure du document. Pour chacun d'eux, nous avons réalisé une étude linguistique et une étude sur corpus qui nous a conduit à proposer, dans la deuxième partie de ce document, un modèle d'analyseur robuste constitué de plusieurs grammaires qui tiennent compte du type de phénomène à traiter lors de l'analyse.

En ce qui concerne les ambiguïtés structurelles, spécialement le rattachement prépositionnel, nous proposons dans la troisième partie de ce mémoire une méthode qui permet d'améliorer leur traitement dans le cadre de l'analyseur syntaxique que nous avons utilisé.

Chapitre 3

Étude de différents phénomènes

3.1 Introduction

Les résultats de l'évaluation de quelques analyseurs existants, sur des corpus variés, font ressortir l'une des difficultés principales à laquelle se heurtent la plupart des systèmes actuels : la large couverture de phénomènes présents dans les corpus est réalisée au détriment de la qualité des analyses. La variation de phénomènes linguistiques a donc d'importantes implications pour l'analyse syntaxique robuste.

Ainsi, les grammaires qui généralisent la description linguistique ont tendance à être imprécises, parce qu'elle ignorent de nombreuses formes linguistiques. D'autre part, des analyseurs très spécifiques (sorte de systèmes *ad hoc*) fournissent des analyses précises pour un certain domaine mais s'avèrent inappropriés à d'autres domaines.

Ces constats nous ont menée à proposer un modèle d'analyseur fondé sur une division de l'analyse linguistique en deux grands niveaux (deuxième partie de la thèse). L'établissement de ces niveaux est motivé par l'étude des phénomènes que nous présentons ici et que l'analyseur doit prendre en compte.

Nous faisons ainsi une première différence entre les phénomènes généralement bien modélisés et ceux qui entraînent des difficultés de traitement (au niveau de leur identification automatique, de leur structure ou des relations qui s'établissent entre les différents éléments les composant). Parmi ces derniers, on retrouve des phénomènes que nous avons brièvement présentés au chapitre précédent, à savoir des phénomènes liés à des marques non lexicales (par exemple la ponctuation) ou lexicales (par exemple le rattachement prépositionnel).

Dans ce chapitre, nous présentons une étude linguistique et empirique de quelques phénomènes que nous avons considérés comme complexes du point de vue de leur traitement syntaxique automatique, et qui feront l'objet de modélisation par des grammaires spécialisées dans notre système (*cf.* chapitre 6) :

- phénomènes liés à la ponctuation : entités entre guillemets ou parenthèses (3.3) ;
- phénomènes liés à la structure du document : listes et énumérations (3.4 et 3.5) ;
- phénomènes liés à la ponctuation et à la structure du document : titres (3.6).

3.2 Précisions terminologiques

Par souci de clarté, nous avons considéré nécessaire de définir au préalable des notions que nous allons employer fréquemment dans la suite de ce document.

*** Phrase**

D'après l'étude sur corpus que nous avons réalisée, la notion classique de *phrase* ne s'avère pas adéquate à notre travail. En effet, il s'agit d'un concept trop *primitif* de découpage d'un texte sur des critères strictement linguistiques. Ainsi, cette notion ne tient pas compte de phénomènes mettant en cause des aspects plus structuraux comme la présentation logique des énoncés ou leur apparence visuelle [Luc00].

Nous avons considéré nécessaire d'élargir cette notion. On définit alors une *phrase* comme suit :

> « Phrase » : entité complète au sein du texte, c'est-à-dire, entité ne pouvant pas être subdivisée en plusieurs suites ayant chacune une fonction particulière dans le texte.

Les éléments qui intègrent une phrase sont liés les uns aux autres par des relations de dépendance qui ne dépassent pas les limites de cette entité.

Par leur nature, nous distinguons trois types de phrases : la phrase au sens « classique », la liste et le titre[1].

Finalement, selon leurs caractéristiques morphosyntaxiques, nous faisons la différence entre des phrases de premier niveau (phrases N1 ou noyau) et les phrases de deuxième niveau (phrases N2 ou intégrant des phénomènes complexes). Nous les définirons en détail dans les chapitres 5 et 6, respectivement.

*** Syntagme noyau**

Dans notre approche, à la base de toute analyse linguistique, nous faisons un découpage en « syntagmes noyau » (*chunks*, [Abn96]). Comme nous l'avons mentionné plus haut (*cf.* chapitre 1) il s'agit d'ensembles non récursifs des syntagmes classiques motivés par une preuve psycholinguistique :

> « Syntagme noyau » : noyau non-récursif d'un constituant (syntagme traditionnel), qui s'étend du début du constituant jusqu'à sa tête et qui n'inclut pas de dépendants au delà de cette tête.

En effet, la phrase *« Les achats de matières premières concernent surtout les industries de transformation. »* serait segmentée comme (1) selon une approche « classique » et comme (2) selon la perspective de découpage en syntagmes noyau :

[1] Pour d'autres traitements linguistiques, d'autres types de phrases peuvent être envisagées. Dans sa thèse, F. Trouilleux découpe les textes en *sections* ou *articles*, qui sont des ensembles de phrases où certains éléments les intégrant peuvent avoir des liens de co-référence entre eux [Tro01].

(1) *[Les achats [de matières premières]] [concernent] surtout [les industries [de transformation]] .*

(2) *[Les achats] [de matières] [premières] [concernent] surtout [les industries] [de [transformation] .*

Lors du développement de la grammaire, nous avons été amenée à utiliser les douze types suivants de syntagmes noyaux :

- NP (syntagme nominal),
- AP (syntagme adjectival),
- PP (syntagme prépositionnel),
- IV (syntagme verbal infinitif),
- FV (syntagme verbal conjugué),
- GV (syntagme verbal participe présent),
- PAP (participe passé),
- SBC (proposition subordonnée),
- BG (début de groupe, conjonction de subordination),
- ADV (adverbe),
- CONJ (conjonction),
- DAT (date).

Par cohérence avec les sorties de notre analyseur, nous allons dorénavant utiliser cette terminologie pour faire référence aux différents types de syntagmes (par exemple, nous utiliserons `NP` pour parler des syntagmes nominaux et non pas `SN`).

*** Patron syntaxique**

Lors de l'analyse de phénomènes liés à la ponctuation (3.3), on utilise cette notion pour étudier leurs caractéristiques morphosyntaxiques.

« Patron syntaxique » : structure composée d'une tête ou noyau syntaxique éventuellement suivie d'une expansion.

Par exemple, le patron syntaxique de *« toutes catégories de produits confondues »* est un syntagme nominal (`NP{toutes catégories}`) suivi d'une expansion composée d'un syntagme prépositionnel et d'un participe passé (`PP{de NP{produits}} PAP{confondues}`).

*** Expansion**

Il s'agit d'un élément linguistique qui dépend d'une tête syntaxique. Le nombre d'éléments pouvant composer une expansion ainsi que leur combinatoire sont élevés.

« Expansion » : syntagme ou un ensemble de syntagmes (avec éventuellement des signes de ponctuation tels qu'une ou des virgule(s), deux points, point-virgule) qui agit en tant que complément de l'élément noyau (syntagme initial après la marque de ponctuation ouvrante).

Dans l'exemple *« voir notre interview en pages 20 et 21 »*, après la marque de ponctuation ouvrante (*«*) et la tête (le verbe à l'infinitif *voir*), l'expansion E est composée d'un syntagme nominal (*« notre interview »*) suivi d'un syntagme prépositionnel contenant une coordination de syntagmes nominaux (*« pages 20 et 21 »*).

Ces précisions étant faites, nous présentons dans la suite de ce chapitre une étude linguistique suivie d'une étude sur corpus de différents phénomènes que nous avons considérés complexes du point de vue de leur traitement syntaxique et pour lesquels nous avons créé des grammaires spécialisées (décrites au chapitre 6).

3.3 Phénomènes liés à la ponctuation

Certains auteurs [SA97] constatent que les traitements spécifiques des marques de ponctuation en linguistique informatique sont peu nombreux, mis à part le travail de quelques auteurs ([Bri94], [Whi95], [Jon94], [Jon96]). Nous nous y intéressons tout d'abord d'un point de vue théorique et plus tard (*cf.* 6.3) du point de vue de l'implémentation d'une grammaire qui traite spécifiquement certaines marques de ponctuation.

3.3.1 Briève étude de la ponctuation

D'après [Jon96], il serait possible de distinguer quatre *entités orthographiques* : les entités lexicales (composées de caractères alphabétiques), les entités numériques (qui peuvent apparaître ensemble ou séparées des entrées lexicales ou de ponctuations spécifiques), les entités graphiques (des images) et la ponctuation. Toute information orthographique qui n'est pas alphanumérique ou graphique est considérée comme ponctuation.

On distingue trois types de phénomènes :

- sub-lexicaux : à l'intérieur d'unités lexicales (point dans les abréviations, tiret, apostrophe) ;
- inter-lexicaux : tout ce qui, conventionnellement, est considéré comme ponctuation ;
- super-lexicaux : phénomènes qui se produisent à un niveau supérieur au lexical, spécialement lors de la représentation textuelle (paragraphes, police, soulignés...)

La figure 3.1 récapitule la classification proposée par [Jon96].

Si on s'intéresse aux phénomènes inter-lexicaux, on peut alors établir une division du point de vue de leur sémantique selon que ces phénomènes sont spécifiques à la source (leur sens peut varier, par exemple #, $, @) ou indépendants de la source (leur sens ne varie pas)[2]. Dans ce cas, ils sont alors *adjonctifs* ou *conjonctifs*.

D'après [Nun90], les *phénomènes adjonctifs* correspondent à des marques qui encadrent des éléments lexicaux (*enclosing marks* selon [Mey87]). Les *phénomènes conjonctifs* (*separators*, *ibid.*) sont des marques qui permettent de séparer les éléments lexicaux.

Notre étude portera dans un premier temps sur quelques phénomènes adjonctifs (que

[2]Ils sont cependant dépendants de la langue.

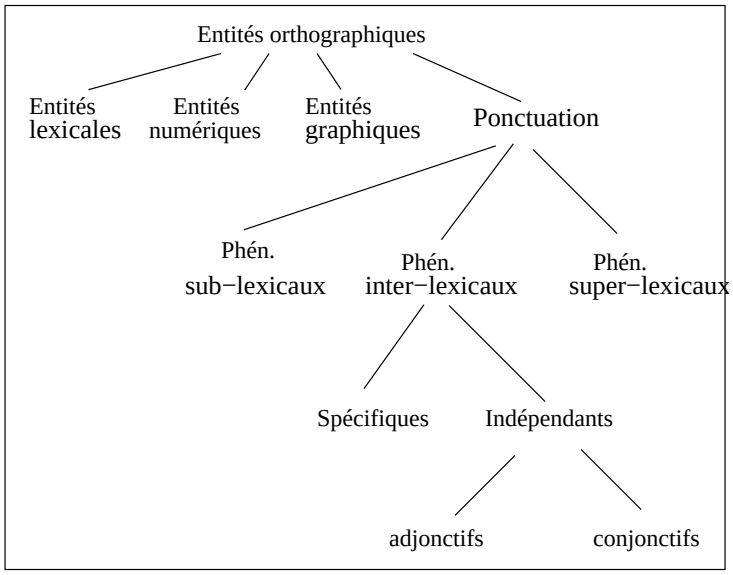

FIG. 3.1 – Hiérarchie des marques de ponctuation.

nous appelons « phénomènes de clôture ») et plus tard sur quelques phénomènes conjonctifs (« phénomènes de séparation »)[3].

Marques et phénomènes « de clôture »

Nous avons choisi de nous focaliser sur deux marques de clôture particulières : les parenthèses et les guillemets[4]. Nous avons considéré les variations concernant les manifestations orthographiques de ces deux marques :

- parenthèses ( ) et crochets [ ], d'une part ;
- guillemets français « » et guillemets anglais “ ”, d'autre part.

Pour les guillemets, nous considérons indifféremment la variante dite anglaise et française (en typographie française). En théorie, ils sont utilisés spécifiquement dans l'une ou l'autre langue mais, en pratique, en français l'emploi des deux types est plutôt aléatoire (selon l'auteur ou le support, on utilisera indifféremment l'un ou l'autre type). Il est à noter que ces marques de clôture sont présentes dans tous les systèmes d'écriture[5].

En tenant compte de ces marques, nous définissons les *phénomènes de clôture* comme des structures syntaxiques circonscrites par des marques du même nom. Voici quelques exemples :

[3]Dans le chapitre 6, nous décrivons les règles de grammaire associées à ces phénomènes.

[4]À l'heure actuelle, nous n'avons pas pris en compte d'autres marques de clôture comme les tirets (− −) ou les clefs ({ }).

[5]En espagnol, par ailleurs, les points d'interrogation (?) et d'exclamation (!) seraient considérés comme des marques de clôture car elles entourent une entité (généralement la phrase).

(3) *En 1994, le Congrès National d'Afrique* **(ANC)**, *dirigé par Nelson Mandela, a été porté au pouvoir à la suite de la première élection multiraciale de l'histoire du pays.*

(4) *La délégation des pouvoirs relève du principe ultime* « **respondeat superior** », *autrement dit, la délégation est contrôlée par la loi elle-même.*

Deux marques de clôture peuvent parfois coexister, partiellement ou totalement :

(5) *Les cinq premiers appelants* **(les « appelants du groupe Adler »)** *sont les parents d'enfants qui fréquentent des externats juifs.*

(6) *La télévision est sa bête noire, de TF1 à la cérémonie des Césars et Hervé Bourges* **(« le roi des courges »)**, *les hommes de droite l'inspirent, les socialistes l'agacent et il a décidé d'adopter un petit communiste (...).*

Le nombre d'éléments lexicaux *à l'intérieur de ces marques* est variable (de 1 à n). Dans nos corpus (décrits en section 3.3.2), nous avons obtenu automatiquement des moyennes de deux mots et démi pour les parenthèses et de presque onze mots pour les guillemets. Le nombre de mots est souvent plus réduit à l'intérieur des parenthèses que des guillemets (voir tables 1 et 2).

Marques et phénomènes « de séparation »

On appelle *phénomènes de séparation* ou *séparateurs* [Nun90] les marques de ponctuation dont la fonction est de délimiter des parties d'unités[6]. Exemples :

(7) *La Reserve Bank of South Africa a fixé quatre conditions pour cela :*
- a. un renforcement des mesures de contrôle des changes ;
- b. le maintien d'une politique budgétaire stricte ;
- (...)

(8) *Dans le cas de la viande bovine, du colza, du blé et des huiles végétales, les contingents tarifaires établis sont de 26.254 tonnes, 871 tonnes, 97.333 tonnes et 61.083 tonnes, respectivement.*

Comme pour les phénomènes de clôture, ces marques sont ambiguës du point de vue de leur fonction ou de la nature des items qu'ils séparent [NBH01].

Voici la liste des séparateurs avec leurs caractéristiques principales :

** Les deux points :*
Ils introduisent des citations, c'est-à-dire du discours rapporté (ils sont alors parfois suivis de guillemets). Ils apparaissent aussi lors de la mise en valeur d'une relation logique (explication, justification, conséquence, opposition...) [DSCH97].

[6]Certaines de leurs fonctions et usages seront étudiés ultérieurement (deux points pour les listes, point-virgule pour les items des listes, virgules pour des énumérations).

** Le point-virgule :*

Il sépare des phrases grammaticalement complètes dont les sens sont liés [DSCH97]. Il dénote une pause de durée moyenne. Il est présent aussi à la fin des alinéas (items) d'une liste.

** La virgule :*

Elle est encore plus ambiguë du point de vue de sa fonction [BSA98]. La virgule sépare des éléments semblables (se substituant à la conjonction). Avec les deux points et le point-virgule, la virgule fait partie des séparateurs secondaires (*secondary boundary marks*, [NBH01]). Ils marquent des limites plus faibles par rapport au point et aux symboles d'interrogation et d'exclamation (*primary boundary marks*).

** Le tiret :*

Il sépare des unités à part dans le discours (incises), mais aussi des items des listes [SA98]. Il est aussi présent dans la transcription de dialogues pour marquer les différentes répliques.

Tous ces séparateurs sont présents, avec un rôle spécifique, au sein d'une même structure : les listes (*cf. 3.4*).

3.3.2 Observations empiriques

Pour l'étude sur corpus de ces phénomènes, nous avons utilisé un ensemble de textes de 122.782 mots provenant de domaines variés : rapport économique, manuel technique, rapports juridiques, rapports scientifiques (médecine et énergie), journaux (*Le Monde*, *Libération*, *La Tribune*, *Les Echos*).

La différence entre le nombre d'occurrences d'un phénomène et le nombre de phrases contenant ce phénomène s'explique par le fait qu'une phrase peut présenter plus d'un couple de parenthèses ou plus d'un couple de guillemets. Cette différence est plus significative pour les parenthèses car souvent, une même phrase contient jusqu'à six ou sept couples de parenthèses ou crochets (cas des phrases avec sigles ou références bibliographiques).

Les parenthèses s'avèrent de loin le phénomène de clôture le plus fréquent dans les différents corpus. La présence de guillemets, quel que soit leur type (“ ” ou « »), est beaucoup moins significative.

Si on tient compte de la longueur des phrases, ce sont les guillemets français (« ») qui apparaissent dans des phrases plus longues (une moyenne de 24 mots par phrase). En général, ce sont aussi les guillemets qui contiennent le plus grand nombre de mots (une moyenne de 11,5 mots). En revanche, la moyenne du nombre de mots dans des parenthèses et crochets est beaucoup plus réduite. L'usage est donc très différent.

Phénomène	*Mots par phrase*	*Mots entre marques*
Parenthèses	13,3 mots	2,9 mots
Crochets	8,8 mots	1,2 mots
Moyenne	**12,3 mots**	**2,5 mots**

Table 1. Nombre de mots (parenthèses).

Phénomène	*Mots par phrase*	*Mots entre marques*
Guillemets “ ”	13,5 mots	10,2 mots
Guillemets « »	24,1 mots	11,5 mots
Moyenne	**16,1 mots**	**10,8 mots**

Table 2. Nombre de mots (guillemets).

Dans les sous-sections suivantes, nous caractérisons de façon générale chaque phénomène et nous montrons sa distribution dans les corpus par rapport à ses patrons syntaxiques (la nature des éléments linguistiques à l'intérieur des marques de clôture). Cette étude de corpus permet de dégager les principales fonctions syntaxiques adaptées aux phénomènes de clôture que nous modélisons plus loin (chapitre 6).

3.3.3 Parenthèses

D'après [Nun90] et du point de vue de leur catégorie syntaxique, les parenthèses sont des « adjoints » (*adjuncts*) à l'intérieur d'une phrase ou d'un paragraphe. Ce sont des constituants allant d'un seul mot à une phrase complète. Ils ne peuvent pas apparaître en position initiale, c'est-à-dire antéposés à la catégorie qui les domine (et de laquelle ils dépendent). Généralement, ils ne peuvent pas contenir un « adjoint » du même type, à moins qu'il s'agisse d'une variante orthographique, c'est-à-dire que *(x (y))* serait rare mais pas (x [y]).

Du point de vue pragmatique, les parenthèses encadrent une indication accessoire. Dans ce sens, ils ont une valeur identique à la double virgule et aux tirets. Les éléments entre parenthèses peuvent être grammaticalement indépendants du reste de la phrase ou lui être reliés [DSCH97].

Nous avons étudié la nature morphosyntaxique des éléments intégrant des parenthèses. Pour ce faire, nous avons calculé leurs « patrons syntaxiques » en utilisant les mêmes corpus décrits en 3.3.2.

La Table 3 montre la distribution des parenthèses dans les corpus (crochets inclus) selon la nature des éléments syntaxiques et/ou lexicaux qui les intègrent. Les abréviations employées sont celles mentionnées plus haut (*cf.* 3.2) ; `Prop` correspond à une proposition de type `SVO`.

Les patrons syntaxiques dans le tableau représentent la tête par un syntagme (`NP`, `AP`, etc.) et l'expansion qui en dépend par le symbole E :

Patron syntaxique	*Nbre d'occurrences*	*% d'apparition*
(NP E)	1358	82,5 %
(AP E)	56	3,4 %
(PP E)	55	3,3 %
(Prop)	27	1,6 %
(IV E)	25	1,5 %
(FV E)	19	1,1 %
(BG/SBC E)	15	0,9 %
(PAP E) ou (ADV E)	14	0,9 %
(CONJ E) ou (DAT E)	11	0,7 %
(GV E)	6	0,4 %
Autre	34	2 %
Total	**1645**	**100 %**

Table 3. Distribution de patrons syntaxiques intégrant des parenthèses.

On constate qu'une grande majorité des parenthèses sont composées de syntagmes nominaux (simples ou multiples, suivis d'autres types d'éléments ou non). Exemples :

(U.E.)
(NP{groupements} PP{de NP{femmes}} , principalement)
(NP{hôtels} , NP{restaurants} , NP{stations} PP{de NP{vacances}})

La catégorie *Autre* correspond à des parenthèses contenant souvent des symboles (+ NP), (< NP PP), (FV < NP), etc.

(NP{Indice} PP{de NP{Karnofsky}} < NP{70})
(+ NP{NOUN{7,2 %}})

Un bon nombre de patrons syntaxiques contiennent des virgules (il s'agit dans la plupart de cas d'énumérations) :

(NP{Société Générale} , NP{CL} , NP{Natexis Capital})
(AP{battantes} , AP{coulissantes} , AP{pliantes} , PP{à NP{tambour}} ...)

Certains patrons contiennent des ponctuations autres que des virgules : deux points, points d'interrogation, etc.

(NP{photo 11} : NP{Mélanome} PP{sur NP{mélanose}} PP{de NP{Dubreuil}})

(IV{sans démarrer}!)

Un nombre réduit d'exemples trouvés contient des guillemets imbriqués dans des parenthèses : (« FV Coord FV »), (NP NP, NP " NP "), etc. ou bien présente des parenthèses (ou crochets) imbriqués.

(« FV{payez} et FV{emportez} »)

(NP{ELF Prestigrade} " NP{TS} " , NP{ELF Multigrade} " NP{S} ")

(NP{Figure} NP{1} (NP{b}))

L'utilisation de crochets est beaucoup moins fréquente que celle des parenthèses. Ils sont employés pour encadrer une information accessoire à l'intérieur des parenthèses ou

bien dans des usages très spécifiques (références bibliographiques, par exemple). Ils n'apparaissent pas dans tous les types de corpus ; leur emploi est restreint (corpus juridique et scientifique, principalement).

[NP{1995}]

[NP{TRADUCTION}]

[NP{L'article} NP{23}]

[NP{KANG1992}]

3.3.4 Guillemets

Comme les parenthèses (bien que [Nun90] n'en parle pas explicitement), les guillemets sont aussi des « adjoints » du point de vue de leur catégorie syntaxique. Ce sont aussi des constituants allant d'un seul mot à une phrase complète (voire plusieurs). À la différence des parenthèses, ils peuvent apparaître en position initiale, c'est-à-dire antéposés à la catégorie qui les domine (et de laquelle ils dépendent), aussi bien que postposés. Exemple :

Les huiles sont de qualités variables et numérotées sous la forme **"xWy"** *(15W40 par exemple).* **"x"** *représente la viscosité à froid.*

Dans cet exemple, *"xWy"* succède la tête syntaxique dont il dépend (*forme*) alors que *"x"* apparaît devant sa tête (*représente*). Indépendamment de la position occupée, les structures entourées de guillemets dépendent toujours d'un autre élément dans la phrase.

D'un point de vue pragmatique, les guillemets marquent les limites d'un texte considéré par l'auteur comme « étranger » à son propre discours [DSCH97]. On constate différents usages, notamment citations, dialogue rapporté et mise en valeur d'un mot (en concurrence avec d'autres 'moyens' comme l'utilisation des italiques, du souligné, etc.). L'utilisation d'un type de guillemets ou autre (français ou anglais) est souvent indifférente (subjective ?) et dépend de l'origine du corpus, de l'auteur...

La Table 4 montre la distribution des guillemets dans le corpus selon la nature des éléments syntaxiques et/ou lexicaux qui les intègrent :

Patron syntaxique	*Nbre d'occurrences*	*% de phrases*
« NP E »	217	48,9 %
« Prop »	52	11,9 %
« AP E »	51	11,7 %
« FV E »	27	6,2 %
« IV E »	27	6,2 %
« PP E »	21	4,8 %
« SBC E »	18	4,2 %
« PAP E »	10	2,3 %
« ADV E »	5	1,1 %
Autre « ... »	7	1,6 %
Total	**435**	**100 %**

Table 4. Distribution de patrons syntaxiques intégrant des guillemets.

Comme pour les parenthèses, ce sont les structures avec des syntagmes nominaux les plus fréquentes (en général des noms seuls ou bien des combinaisons de NPs, APs et PPs). Il faut remarquer la présence plus fréquente de phrases complètes, qu'elles soient noyau ou pas. Parfois, aussi, plus d'une phrase peut être comprise dans des guillemets.

« NP{subterfuge} »

« NP{Convention} AP{européenne} PP{de NP{sauvegarde}} PP{des NP{droits de l'homme}} et PP{des NP{libertés}} AP{fondamentales} »

"{NP{Il} FV{se trompait} {complètement} .} "

Comme pour les parenthèses, aussi, on constate la présence d'autres signes de ponctuation à l'intérieur des guillemets, généralement deux points, des points d'exclamation, interrogation ou des parenthèses. Nous n'avons pas trouvé de guillemets imbriqués dans les corpus utilisés.

" NP{Guide} PP{des NP{débouchés}} AP{commerciaux} PP{en NP{Afrique du Sud}} : AP{Additifs} et NP{ingrédients} AP{alimentaires} "

" NP{Un AP{premier} ministre} PP{de NP{droite}} ou PP{de NP{gauche}} ? "

« IV{faire monter} NP{les coûts} et , PP{par NP{ce biais}} , [de] IV{pousser} NP{les prix} PP{à NP{la hausse}} »

3.4 Listes

Les listes, présentes de façon variable selon le type, le genre ou le domaine du corpus, sont des phénomènes qui, de par leurs caractéristiques, mettent en jeu des paramètres liés à la structure du document (espaces verticaux et horizontaux). Nous avons jugé indispensable de prendre en compte ces paramètres lors de la modélisation des listes dans le cadre de la grammaire de notre analyseur car ils sont à la base de la définition des listes.

La notion de « liste » doit cependant être bien définie. Certains auteurs [LGDM+99] appellent *énumération* tout type de structure énumérative. Pour nous, les *énumérations* sont des structures énumératives où les différents items ne présentent pas des paramètres dispositionnels (*cf.* section 3.5). Nous considérons alors les *listes* comme des structures énumératives organisées verticalement (*vertical lists* [Whi95]).

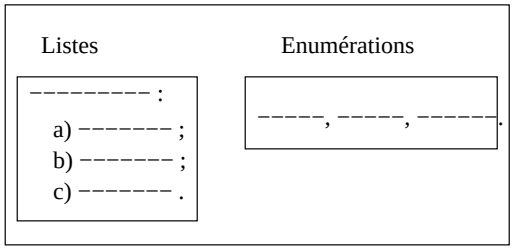

FIG. 3.2 – Typologie de structures énumératives.

D'après [LGDM+99], les « énumérations » selon leur dénomination (**listes** pour nous) « sont le plus communément appréhendées comme un moyen de mettre en relief de façon identique des objets ou des entités fonctionnellement équivalentes ». Cette mise en relief des items est due précisément à l'organisation verticale : chaque élément représentant une entité énumérée (*item*) apparaît disposé en dessous du précédent, avec un *séparateur* ou *organisateur* devant et, dans la plupart des cas, un séparateur derrière.

Nous allons par la suite caractériser du point de vue linguistique ces séparateurs puis les deux composants principaux des listes : *l'amorce* et l'ensemble d'items. Finalement, nous montrons quelques relations sémantiques établies au sein des listes.

3.4.1 Les marques

Les listes présentent un nombre important de marques de différentes sortes. La notion de *marque* doit être comprise au sens large : elle inclut de la typographie, des paramètres de mise en forme, de la ponctuation, etc. D'après [LGDM+99] on peut classifier ces marques en trois groupes :

a) marques typographiques

b) marques dispositionnelles

c) marques lexico-syntaxiques

a) Marques typographiques

Les marques typographiques sont fondamentales dans les listes. Elles aident à une meilleure visualisation des différents composants de cette structure énumérative. Il en existe différents types, chacun ayant une position et une fonction particulière au sein de la liste.

En fin d'amorce on retrouve la plupart des cas les deux points [**:**]. Ils introduisent l'ensemble d'items de la liste.

En début d'item, on distingue deux types de marques typographiques :

- *séparateurs*, des marques de ponctuation telles que [**-**], [*****] et [**.**] ;
- *organisateurs*, des symboles alphanumériques suivis de parenthèse fermante, point, tiret, etc. comme [**a)**], [**i.**] ou [**1-**].

En fin d'item, les séparateurs plus usuels sont le point-virgule [*;*], la virgule [*,*] et le point [.], bien que dans certains cas il n'y ait pas du tout de marque.

Finalement, en fin de liste c'est le point [**.**] le séparateur le plus utilisé (il équivaut au point de fin de phrase).

b) Marques dispositionnelles

Comme les marqueurs typographiques, les marques dispositionnelles sont importantes pour la mise en page (*layout*) : elles aident au repérage visuel de la liste. On distingue les *blancs* [LGDM+99] (ou *espaces* [DH96]) verticaux et horizontaux. Les premiers sont des

retour(s) à la ligne entre l'amorce et le premier item et parfois entre les différents items ; les espaces horizontaux sont des tabulations en début d'item.

c) Marques lexico-syntaxiques

Les marques lexico-syntaxiques se trouvent dans l'amorce. Il s'agit de :

- numéraux, groupes nominaux spécifiques (« trois », « les ... suivants », « plusieurs », « certains », etc.)
- classifieurs *génériques* (« critères », « possibilités », « contraintes », « étapes », « types », etc.)
- classifieurs de discours (« remarques », « réponses », « observations », etc.)

Le deuxième groupe est sous-divisé par [LGDM+99] en deux groupes *classifieurs génériques* et *classifieurs de domaine*, ces derniers étant des noms moins généraux et particuliers à un domaine (« critères », « contraintes »). Cette distinction ne nous semble pas nécessaire dans le cadre de l'analyse syntaxique robuste[7].

3.4.2 L'amorce

L'amorce ou séquence introductrice (ou encore *déclencheur* [Bus00]) est une phrase ou ensemble de segments (groupes nominaux, adjectivaux, prépositionnels, etc.) qui a comme fonction d'introduire les items de la liste. Elle est composée principalement d'un *noyau*, elle peut contenir des éléments *"annexes"* (parfois d'autres phrases) et elle finit, en général, par deux points.

Dans la plupart des cas, le *noyau de l'amorce* contient un mot qui agit en tant qu'hyperonyme de l'ensemble d'items de la liste (appelé aussi *classificateur* [Bus00]) et un ou plusieurs *modifieurs* de celui-ci. Les modifieurs apportent des précisions, c'est-à-dire qu'ils restreignent le champ sémantique du classificateur.

Les *annexes* introduisent ou apportent un complément d'information aux éléments principaux de la liste (noyau et items). Du point de vue syntaxique, les annexes apparaissent seulement dans le cas où le noyau est le complément d'objet direct de la phrase-amorce.

La figure suivante schématise les composants d'une amorce :

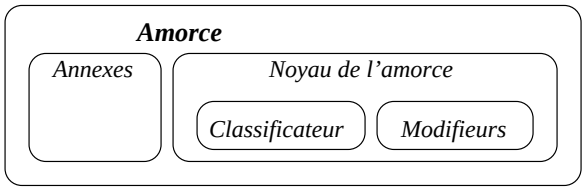

FIG. 3.3 – Éléments composant l'amorce d'une liste.

D'après cette figure, on peut décomposer l'exemple comme suit :

[7] Lors d'une analyse sémantique elle est peut-être importante.

(9) *En procédant de la sorte, nous avons pu mettre en évidence les ingrédients nécessaires à la production d'un catalyseur pour la réduction de l'oxygène en PEFCs :*
- *un métal de transition.*
- *une source d'azote.*
- *une source de carbone.*
- *un traitement thermique.*

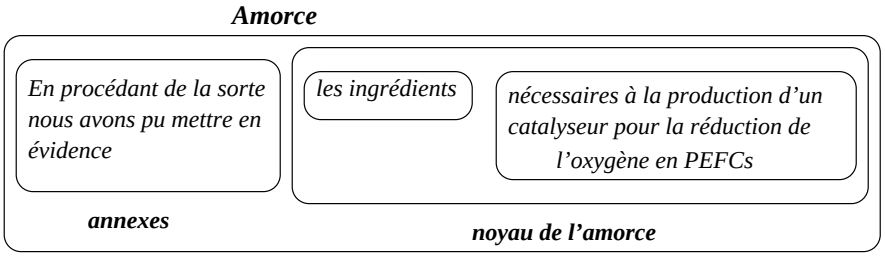

FIG. 3.4 – Amorce avec annexes.

Les annexes n'apparaissent pas toujours dans la structure d'une amorce. Ainsi, elles sont inexistantes lorsque le noyau de l'amorce est syntaxiquement le sujet de la phrase :

(10) *Les différents paramètres favorisant l'interférence avec les systèmes électriques sont :*
- *Faible voltage de circuit,*
- *Faible résistance,*
- *Situation à l'intérieur du compartiment moteur,*
- *Non écrantement par du métal.*

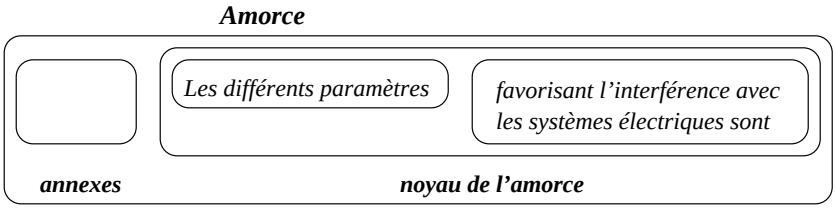

FIG. 3.5 – Amorce sans annexes.

Les exemples précédents permettent d'identifier, toujours du point de vue syntaxique, deux types d'amorces : les *amorces saturées* et les *amorces non saturées*.

A) Amorce saturée

Les amorces saturées (AS) sont des phrases complètes, presque syntaxiquement indépendantes par rapport à l'ensemble d'items de la liste. Exemple :

(11) *Ainsi, si l'on prend les activités les plus importantes qui caractérisent l'industrie de la chaussure, on se rend compte de trois choses essentielles :*
- une faible performance au niveau de la tâche
- le manque d'intégration de chaque activité dans un processus global
- l'absence de polyvalence.

B) Amorce non saturée

Les amorces non saturées (ANS) sont des propositions syntaxiquement incomplètes dont les constituants manquants sont fournis par les items de la liste. Exemple :

(12) *Les dispositions de cette Convention ne seront pas applicables aux personnes dont on aura des raisons sérieuses de penser :*
a) Qu'elles ont commis un crime contre la paix, un crime de guerre ou un crime contre l'humanité (...) ;
b) Qu'elles ont commis un crime grave de droit commun en dehors du pays d'accueil avant d'y être admises comme réfugiés ;
c) Qu'elles se sont rendues coupables d'agissements contraires aux buts et aux principes des Nations Unies.

Un cas précis d'amorce non saturée est celui des *formes réduites*. Dans ces cas, l'amorce est uniquement composée d'un groupe nominal (on peut la considérer comme une proposition avec un verbe *être* elliptique). Exemple :

(13) *Tumeurs primitives bénignes du rachis*
- Granulome éosinophile
- Kyste anévrismal
- Hémangiome vertébral

L'exemple ci-dessus pourrait se paraphraser par *"Les tumeurs primitives bénignes du rachis sont : le granulome, le kyste ..."*. On remarquera que dans ce cas l'amorce est une sorte de titre et pour cette raison elle ne présente pas les deux points habituels à la fin.

Finalement, il existe des amorces où l'annexe et surtout le noyau se trouvent dans la phrase précédente. La recherche de l'hyperonyme noyau doit avoir recours à des techniques de résolution d'anaphore. Dans l'exemple, *« il s'agit »* devrait pouvoir être en relation avec *« exigences »*, qui dans ce cas est l'hyperonyme de *« qualité »*, *« complexité »* et *« durée »*.

(14) *Concernant l'entreprise de céramiques sanitaires, trois exigences essentielles, inhérentes au procès, déterminent fortement l'obtention de la qualité du produit. Il s'agit de :*
- la qualité de fabrication, inhérente au cahier des charges. Elle porte notamment sur les aspects de surface (grain incrusté, fente, couleur...) ;
- la forte complexité. La production est par exemple dépendante des données climatiques dans la mesure où le Mistral génère un air sec qui modifie les temps de prise et de raffermissement ;
- les durées techniques déterminantes.

3.4.3 L'ensemble d'items

La disposition verticale de l'ensemble d'items est la caractéristique visuelle principale des listes. Ces items constituent les différents éléments énumérés et dans la plupart des cas ils présentent une structure parallèle (i.e. des noms suivis de groupes prépositionnels, des propositions, etc.)[8], c'est-à-dire qu'ils sont équivalents morpho-syntaxiquement. En effet, les différents items d'une liste ont la même fonction syntaxique et dépendent de la même tête.

Chaque item de l'ensemble est introduit par une marque typographique. Le cas où les débuts d'items n'ont pas de marque de séparation n'est pas fréquent.

Pour les amorces saturées (AS), l'ensemble d'items constitue une liste d'éléments hyponymes du classificateur. Exemple :

(15) *Il existe deux familles de* **moteurs automobiles** *:*
1) Les **moteurs à explosion**, *dans lesquels la combustion de l'essence est amorcée par l'étincelle de la bougie ;*
2) Les **moteurs DIESEL**, *dont la combustion est amorcée par l'injection de gas-oil sous pression dans de l'air comprimé, ce qui produit une auto-inflammation.*

Dans le cas des amorces non saturées (ANS), l'ensemble d'items est constitué, la plupart des cas, des objets d'un verbe transitif (*inclure, recommander, comprendre, présenter,* etc.), des attributs du verbe *être*, des compléments prépositionnels d'un verbe intransitif (*porter sur, dépendre de, viser à, consister à,* etc.). Exemples :

(16) *Sur 202 salariés, il y a :*
. 70 manoeuvres (dont 69 femmes)
. 44 OS1 (dont 30 femmes)
. 72 OS2 (hommes)
. 16 mécaniciens (hommes)

(17) *Il s'agit de développer une dynamique interne qui vise à :*
- disposer d'un outil de production le plus performant possible (innovation technologique - productivité - qualité)
- maîtriser les nouvelles technologies (plus la technologie est performante plus on arrive à obtenir la complexité croissante du produit demandée)
- développer la créativité et l'adapter aux besoins des marchés (existants et à créer)

Le nombre d'items est assez variable, le minimum étant deux, le maximum autour de la dizaine, avec une moyenne de trois ou quatre items par liste.

Du point de vue morphosyntaxique (nature des items), un grand nombre d'items est constitué d'un syntagme nominal suivi ou non d'une *expansion*[9].

[8] Il est plutôt rare que les items d'une même liste aient des structures morphosyntaxiques très différentes.

[9] Rappel : l'*expansion* est l'ensemble de syntagmes modifiant le syntagme principal ou noyau.

3.4.4 Relations sémantiques

Du point de vue sémantique, on peut distinguer deux types de relation entre le noyau de l'amorce et les items de sa liste.

Le premier cas est celui d'une relation d'hyperonymie-hyponymie (relation « *is a* »), c'est-à-dire que le noyau de l'amorce a une signification plus générale que les items, qui sont plus spécifiques. C'est le cas de l'exemple suivant :

(18) *Il existe plusieurs appellations pour les* **moteurs automobiles** *:*

Moteur en ligne *: les cylindres, pistons, bielles, et vilebrequin sont monté sur le même axe, l'un à la suite de l'autre.*

Moteur en V *: les cylindres ont une disposition opposée qui varie de 15 à 90 avec un axe identique (vilebrequin).*

Moteur à plat *: les cylindres sont opposés deux à deux, ce qui équivaut à un moteur en V et l'angle est de 180.*

L'autre type de relation sémantique est une relation d'holonymie-méronymie (relation « fait partie de »). Dans ce cas, le noyau représente un "tout" et les différents items ses parties. Chaque item est ainsi un composant (du point de vue sémantique) de l'amorce :

(19) *Description d'un* **pneumatique** *:*

1) La **carcasse** *est le squelette du pneu, elle est composée de plusieurs nappes (...).*

2) Le **talon** *est l'élément de liaison avec la jante, il se compose d'une tringle en acier sans soudure (...).*

3) Les **flancs** *encaissent la charge, il protègent la carcasse des chocs latéraux et assurent la souplesse du pneumatique.*

4) La **bande** *de roulement est la partie en contact avec le sol.*

La relation d'holonymie-méronymie est moins fréquente que la relation d'hyperonymie-hyponymie : sur cinquante listes extraites au hasard des corpus variés, après vérification manuelle, 36 listes (72 %) sont des cas d'hyperonymie-hyponymie, 3 listes (6 %) des cas d'holonymie-méronymie, le reste étant des cas n'appartenant à aucune de ces deux possibilités.

3.4.5 Étude sur corpus

Les corpus utilisés pour cette étude sont des corpus variés récupérés sur le Web (mai 2001), moyennant des requêtes ciblées[10]. Ils proviennent de quatre domaines différents : juridique, économique, technique, scientifique (médical et physique).

De ces corpus, nous avons recupéré 217 listes. Elles présentent les caractéristiques suivantes :

[10]Nous avons soumis des requêtes au moteur de recherche avancée d'`Altavista` telles que `suivants :`, `comme NEAR :`, `sont NEAR :`, etc.

Corpus	*Nbre listes*	*Moy. mots/liste*	*Moy. mots/amorce*	*Moy. mots/item*
Scientifique	72	**59**	12	**10**
Économique	72	65	14	12
Technique	38	63	**11**	13
Juridique	35	**114**	**17**	**20**
total	217	71	14	13

Table 5. Distribution et caractérisation des listes étudiées.

Les listes plus longues sont celles appartenant au domaine juridique (les items incluent souvent plusieurs phrases) ; les listes plus courtes sont celles du domaine scientifique (les items sont souvent des noms ou des syntagmes nominaux courts —deux ou trois mots).

Les plus longues amorces (plus de 17 mots) ainsi que les items plus longs (plus de 20 mots par item) se trouvent dans les corpus juridiques. En revanche, les amorces plus courtes se trouvent dans les corpus techniques et scientifiques (moins de 12 mots). Les corpus scientifiques présentent aussi les items les plus courts (en moyenne 10 mots).

Marques de ponctuation

Typologie de séparateurs présents dans les listes étudiées :

Séparateurs	*Fin d'amorce*	*Fin d'item*	*Fin de liste*
[**:**]	**209**	-	-
[**;**]	-	67	-
[**,**]	2	13	-
[**.**]	2	64	**187**
[ ]	4	**70**	25
autres	-	3	5

Table 6. Distribution des séparateurs selon localisation dans les listes.

La catégorie *autres* inclut [!], [?], ["], [...]. Par exemple :

(20) *En même temps, ils se posent un certain nombre de questions, qui restent sans réponse, ce qui montre bien, entre autres, que la circulation de l'information est faible, mais aussi le partage de la stratégie en matière d'organisation de production. Par exemple :*

- Pourquoi un carrousel sur tel module et pas sur l'autre ?
- A-t-on fait des analyses comparées : productivité-souplesse ? Quelles sont les conclusions ? ... Quelle décision ? ...
- Quelles consignes pour le nouveau poste prévu pour faire la navette : piqûre / réparation / coupe ?, etc.

La fin des items est assez variable, même s'il existe une règle typographique qui conseille de finir les items avec des points-virgules (sauf le dernier qui doit finir par un point).

Le tableau suivant montre une typologie de séparateurs et organisateurs en début d'items des listes.

Séparateurs	*Début d'items*
[**-**]	123
[**.**]	12
[*****]	2
total sép.	**137**
Organisateurs	*Début d'items*
[**1) 1. 1-**]	33
[**a) a. a-**]	25
[**i) i. i-**]	4
total org.	**62**
total autres marques	**18**
total global	**217**

Table 7. Distribution des marques de séparation.

La présence de séparateurs et organisateurs est assez variable, avec une prédominance des marques typographiques sur les alphanumériques (environ deux tiers / un tiers).

Amorces

Concernant les amorces on constate une répartition équilibrée entre amorces saturées (AS) et amorces non saturées (ANS) : 52,5 % des amorces correspondent à des AS et 47,5 % à des ANS.

Corpus	*Saturées*	*Non Saturées*
Scientifique	34	40
Économique	42	29
Technique	26	11
Juridique	12	23
total	114	103

Table 8. Distribution des types d'amorce.

Les amorces saturées prédominent dans les corpus économique et technique ; les non saturées plutôt dans les corpus scientifique et juridique.

Par ailleurs, parmi les ANS il y a 15 formes réduites (dont 13 dans des corpus scientifiques et 2 dans le corpus techniques). Elles représentent 7 % du total des amorces des listes du corpus.

Items

Le tableau suivant montre le nombre d'items par liste classés selon le domaine du corpus. Tel que nous l'avons défini plus haut, une liste présente un minimum de deux

items.

Corpus	*2 items*	*3 items*	*4 items*	*5 items*	*6 items*	*+6 items*
Scientifique	22	**25**	14	5	3	3
Économique	15	20	**21**	4	4	8
Technique	10	**14**	7	2	2	3
Juridique	**13**	9	4	4	2	3
total	60	68	46	15	11	17

Table 9. Nombre d'items des listes par domaine du corpus.

En général, les listes ont de 3 à 4 items (moyenne 3,6 items) ; il n'y a pas beaucoup de différence par domaines, si ce n'est que les corpus juridiques ont souvent des listes avec 2 éléments mais peuvent en avoir aussi jusqu'à 10, alors que dans les autres domaines le maximum est rarement supérieur à 8 items par liste.

Les items sont des patrons morphosyntaxiques se trouvant après les séparateurs/organisateurs (voir la notion d'*Expansion*, 3.2). Exemple :

(21) *Le syndrome sous-lésionnel à la phase d'état comprend des troubles moteurs avec :*
- *une paralysie plus ou moins importante*
- *une exagération des réflexes qui sont vifs*
- *une inversion du réflexe cutané plantaire (signe de Babinski).*

Pour les items de cette liste, les têtes sont les noms (*paralysie, exagération...*) et les expansions leurs modifieurs (*plus ou moins importante, des réflexes qui sont vifs...*).

Patron	*Pourcentage*
NP Expansion	60,4 %
Phrase Expansion	14,1 %
PP Expansion	5,0 %
IV Expansion	4,2 %
AP Expansion	2,7 %
FV Expansion	2,1 %
Autre	11,5 %
total	100 %

Table 10. Patrons morphosyntaxiques des items des listes.

Les syntagmes nominaux suivis d'une expansion sont les composants majoritaires des items. Les phrases complètes sont aussi fréquentes.

3.5 Énumérations

Les énumérations sont un cas particulier de structure énumérative (*cf.* figure 3.2) : elles n'ont pas de structure visuelle (retours à la ligne, blancs verticaux et/ou horizontaux)

mais les différents items énumérés sont équivalents à ceux des listes du point de vue de leur fonction syntaxique.

Étant donné ces différences structurelles, nous avons choisi de les étudier séparément (puis de les modéliser au sein d'une grammaire particulière).

3.5.1 Caractéristiques

On définit une énumération comme une succession de syntagmes principalement nominaux. Une analyse des énumérations en corpus (*cf.* 3.5.3) permet de recenser différentes constructions possibles, avec des marqueurs initiaux et finals variés (nous utilisons les crochets pour marquer les limites droite et gauche des énumérations en exemple).

Les énumérations peuvent apparaître au milieu ou en fin de phrase, mais rarement en début (aucun cas n'a pas été trouvé). Elles peuvent aussi faire partie de listes et titres :

(22) *En Ontario , toutefois (...) on peut dire que la protection des zones présentant un intérêt environnemental comme [les terres humides , les boisés , les ravins et les meilleures terres agricoles] a , à toutes fins utiles , été éliminée.*

(23) *Elle s'altère suite aux contraintes qu'elle subit [(température, frictions, contamination...)] et doit donc être remplacée régulièrement.*

(24) *En demi-finale , vendredi 29 janvier , le Suédois a battu l' Américain Pete Sampras [7 - 6 , 6 - 3 , 7 - 6.]*

(25) *La hausse du dollar a bénéficié aux grandes valeurs exportatrices comme [Unilever, Royal Dutch Shell ou Akzo Nobel.]*

(26) *Il est possible d'obtenir la donnée sous l'une des formes suivantes pour lesquelles est indiqué le volume de données acquis par jour :*
- la donnée moyenne sur les 20 fenêtres,
- une information [(temps, amplitude)] concernant les pics d'arrivée,
- la même information moyenne sur les 20 fenêtres.

(27) *Fruits et fruits à coque [(noix , amandes ...)]*

D'après ces exemples et selon la description de [Mor99], nous pouvons faire les constats suivants :

- une énumération est constituée de deux éléments (ou plus) équivalents du point de vue de la fonction qu'ils ont dans l'unité linguistique dans laquelle ils se trouvent ;

- les items énumérés sont précédés d'un marqueur initial qui peut être typographique (deux points, parenthèse ouvrante) ou discursif (*notamment, tels que, par exemple, comme ...*) ;

- à la fin de l'énumération il y a un marqueur typographique (parenthèse fermante, point, points de suspension) ou discursif (*etc.*).

3.5.2 Typologie

Nous avons caractérisé les énumérations au moyen de deux notions. La première, plus syntaxique, est fondée sur la présence ou l'absence d'une conjonction de coordination entre l'avant-dernier et le dernier item énuméré (principalement *et*, *ou*, *mais*), exemple (28). On parle ainsi d'énumérations *avec ou sans coordination.*

(28) *Le réglage s'effectue dans l'ordre numérique des cylindres : [1, 2, 3 et 4.]*

La deuxième notion est d'ordre plus sémantique. On parle alors d'énumération *ouverte* si la liste d'éléments énumérés n'est pas finie, exemple (29) –elle finit par des points de suspension ou par l'abréviation *etc.*.

(29) *Eclairage : vérifiez le bon fonctionnement [des feux , clignotants , feux de stop , etc .]*

Les énumérations *fermées*, en revanche, sont celles avec un nombre d'items fini, exemple (30).

(30) *À cause de différents facteurs , comme [la santé animale , la dimension culturelle et les habitudes de consommation] , ce secteur est peut-être le plus délicat qui soit.*

Le tableau suivant synthétise les deux notions proposées :

Patrons	*Avec coordin.*	*Ouverte*
a, b, c ...	-	+
a, b, c .	-	-
a, b et c	+	-

Table 11. Récapitulatif typologie des énumérations.

Il n'existe pas d'énumérations *ouvertes* avec items *coordonnés* car ce sont deux notions incompatibles (la présence de la coordination « ferme » en quelque sorte la liste d'items).

3.5.3 Observations à partir de l'étude sur corpus

Le corpus utilisé est un corpus de 151.000 mots provenant de différents domaines et extrait du Web (mars 2001). La moyenne de mots des phrases contenant des énumérations est d'environ 32 mots. La moyenne de mots par énumération est de 10.

Le tableau suivant montre la distribution des énumérations par type de corpus.

Corpus	*Nbre phr*	*Nbre énms*	*Moy. m/phr*	*Moy. m/énm*	*Milieu phr*	*Fin phr*
Journ.	25	26	33,5	10,8	7	19
Techn.	21	23	21,8	7,9	9	14
Éconm.	19	22	30,4	8,7	9	13
Scient.	11	13	42	13,4	7	6
Jurid.	7	7	42,2	15,6	2	5
total	83	91	31,7	10,3	35	56

Table 12. Distribution et position des énumérations étudiées.

La moyenne de mots par énumération (sans compter les signes de ponctuation) est d'environ 10 mots, avec une moyenne de presque 8 mots pour les corpus techniques et de presque 16 pour les corpus juridiques.

Des 83 phrases, il y a 3 listes, 1 titre et 79 phrases « classiques » (ni titres ni listes), dont 8 qui contiennent deux énumérations comme c'est le cas des exemples suivants (le deuxième concerne une énumération imbriquée).

(31) *Les plus grands fournisseurs de volaille vivante [(incluant les poules à bouillir, le canard sauvage, les oies, etc...)] en 1994 étaient [les États-Unis (299 000 $US), le Royaume-Uni (200 000 $US) et les Pays Bas (67 000 $US)].*

(32) *Il y a certains aliments demandés par l'industrie sud-africaine qui ne sont pas produits au pays, notamment [les figues séchées, des fruits comme [les cerises noires et les baies], et les haricots secs].*

Les deux tableaux suivants (13 et 14) présentent la distribution des marqueurs dans les énumérations étudiées. Il est à noter que pour le premier tableau, la catégorie *« Autre »* comprend des marqueurs comme *tels que*, *suivants*, *voici*, etc. :

Corpus	*(*	*:*	*comme*	*notamment*	*ø*	*« Autre »*
Journalistique	4	5	2	2	10	3
Technique	11	5	-	-	4	1
économique	4	6	7	3	2	1
Scientifique	6	3	1	-	2	1
Juridique	-	2	1	1	3	1
total	25	22	11	6	20	7

Table 13. Marqueurs initiaux.

Dans certains cas, deux marqueurs initiaux peuvent coexister. Dans ces cas, on a compté seulement le dernier (par exemple, les deux points pour *tels que :*).

On note l'utilisation de beaucoup plus de marqueurs typographiques (particulièrement des parenthèses, par exemple dans les corpus techniques) et en général peu de marqueurs discursifs. Aussi, on constate une absence importante de marqueurs, quel que soit leur type, dans les corpus journalistiques. Par exemple :

(33) *Peu à peu, s'infiltrent [des déhanchements, des balancements, de faux déséquilibres, des tourbillons.]*

Pour les marqueurs de fin d'énumération, la catégorie *« Autre »* comprend la parenthèse fermante, la virgule et le point-virgule.

Corpus	.	...	*etc.*	*« Autre »*
Journalistique	18	3	1	4
Technique	8	4	9	2
Économique	11	1	2	8
Scientifique	4	2	2	5
Juridique	4	-	1	2
total	45	10	15	21

Table 14. Marqueurs finals.

Comme pour les marqueurs initiaux, dans certains cas deux marqueurs finals coexistent. On a compté alors seulement le premier (la parenthèse fermante pour *) .* , les points de suspension pour *...)*, l'abréviation etc pour *etc...*).

Le point final est de loin le plus important car dans notre corpus il y avait plus d'énumérations en position de fin de phrase.

Finalement, le dernier tableau caractérise les types d'énumérations selon les notions proposées plus haut. La dernière colonne montre la moyenne d'éléments énumérés par type de corpus.

Corpus	*Coordin.*	*Non coordin.*	*Ouvertes*	*Fermées*	*Moy. items*
Journalistique	8	18	4	22	4 (3,6)
Technique	5	18	13	10	3 (2,7)
Économique	18	4	3	19	3 (2,8)
Scientifique	6	7	3	10	3 (2,8)
Juridique	6	1	1	6	4 (4,3)
total	43	48	24	67	3 (3,1)

Table 15. Caractérisation des énumérations.

D'aprés ces données, on constate que la plupart des énumérations de notre corpus sont fermées (67 sur 91, soit 73,4 %) avec une moyenne générale de trois items. La distinction entre énumérations avec ou sans coordination est moins importante (47,3 % et 52,7 %, respectivement).

3.6 Titres

Nous étudions ici un type particulier de structure présentant des caractéristiques linguistiques et structurelles qui la rendent différente des phrases « classiques » et aussi des listes.

3.6.1 Étude linguistique

Les titres sont en effet des *phrases* proprement dites (au sens de notre définition en section 3.2)[11] qui finissent (généralement) sans signe de ponctuation et qui sont séparés

[11] Phrase : entité complète au sein du texte, c'est-à-dire, qu'elle ne peut pas être subdivisée en plusieurs suites ayant chacune une fonction particulière dans le texte.

du reste du texte par des espaces verticaux plus grands que ceux qui séparent deux paragraphes standard. Leur fonction est d'énoncer le contenu de la partie ou sous-partie de texte qui suit. Exemples :

(34) *COURROIE DU VENTILATEUR*

(35) *Remarque*

(36) *Loi n 78−17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*

Dans certains cas, les titres présentent une ponctuation finale : il s'agit de ponctuation forte avec une fonction spécifique au delà du simple marquage de fin de phrase (introduction d'une liste [:], marque interrogative [?], etc.).

(37) *LES COTES DE REGLAGES DU 350cc STANDARD :*

(38) *3.10 La théorie rend-elle compte d'autres interférences rapportées ?*

Les cas de titres avec ponctuation classique de fin de phrase ([.], [;], etc.) correspondent à des usages non normalisés.

Les aspects typographiques des titres sont à remarquer. En effet, la police utilisée peut être indistinctement toute en majuscule ou toute en minuscule (sauf majuscule initiale). Des caractères plus grands que la police du corps du texte sont fréquents, ainsi qu'une marque de relief (gras, cursive, etc.).

Dans certains corpus (manuels techniques, rapports scientifiques, etc.), les titres sont parfois précédés d'organisateurs (chiffres) qui marquent l'ordre logique d'apparition du titre et du texte qui le suit, par rapport à l'ensemble du texte .

Du point de vue morphosyntaxique, dans la plupart des cas, les titres sont des syntagmes nominaux coordonnés ou pas avec des *expansions*. Il peut aussi s'agir de propositions. C'est seulement dans certains types de corpus (par exemple dans les journaux) que les titres sont des phrases au sens classique (elles sont souvent simplifiées ou appauvries).

3.6.2 Étude sur corpus

Pour l'analyse des titres, nous avons utilisé cinq corpus récupérés du Web entre juin et novembre 01, que nous avons classés par domaine (total de 103.377 mots)[12] selon la source. L'ensemble de ce corpus présente plus de titres (20 % des phrases) que dans la moyenne des corpus utilisés jusqu'à présent (environ 15 %)[13].

Le tableau suivant montre les caractéristiques des phrases des corpus et plus précisément des titres (la quatrième colonne montre le pourcentage de titres par rapport au nombre de phrases de chaque corpus).

[12]Domaines du corpus : économique (36.203 mots), juridique (22.894 mots), technique (21.169 mots), scientifique (18.392 mots), journalistique (4.718 mots).

[13]Les comptages que nous decrivons ont été faits automatiquement par des scripts *perl*.

Corpus	*Total phrases*	*Titres*	*Pourcentage*	*Moyenne mots*
Juridique	796	270	33,9 %	4 (3,6)
Technique	751	179	23,8 %	5 (4,8)
Scientifique	925	171	18,4 %	5 (4,9)
Économique	1095	126	11,5 %	7 (6,7)
Journalistique	246	50	20,3 %	3 (3,5)
total	3813	796	20,9 %	5 (4,6)

Table 15. Distribution et moyenne de mots des titres étudiés.

En général, les phrases de ces corpus ont une moyenne de 35,4 mots. Les phrases plus courtes se trouvent dans les corpus scientifiques (15 mots) et les plus longues dans les corpus juridiques (52 mots).

Pour les titres, la moyenne de mots est autour de 5 mots par titre, alors que dans les corpus journalistiques la moyenne est d'environ 3 (beaucoup de titres avec un seul mot et aussi des phrases courtes) et dans les rapports économiques d'environ 7 (beaucoup de syntagmes nominaux avec des *expansions* nominales –via des prépositions– et adjectivales).

Voici quelques exemples, les trois premiers correspondant à des corpus journalistiques, les trois derniers à des corpus économiques :

(39) *Langue de bois*

(40) *Chirac en conseiller de Bush*

(41) *Organisation du travail et contrôle qualité*

(42) *1. 8- L'implication des salariés dans le processus de production*

(43) *B. Rapports sur les possibilités d'exportation de produits*

La table suivante schématise les patrons syntaxiques des titres. La première colonne contient des titres avec des verbes conjugués `FV` (phrases entières, pas de propositions subordonnées). La deuxième colonne correspond aux titres qui commencent par un `NP` et sont suivis d'une *expansion* qui ne contient pas de verbe conjugué.

La troisième colonne est celle des titres avec séparateur initial (organisateur logique). Les colonnes quatre, cinq et six correspondent à des titres qui commencent respectivement par syntagme prépositionnel, syntagme adjectival et verbe à l'infinitif. La dernière colonne montre les titres qui commencent par guillemets, parenthèses ou conjonctions (catégorie *autre*).

Corpus	`FV`	`NP`	`SP`	`PP`	`AP`	`IV`	*autre*
Juridique	22,5 %	**50,7 %**	5,6 %	1,5 %	12,7 %	1,4 %	5,6 %
Technique	11,1 %	35,2 %	**42,6 %**	-	3,7 %	3,7 %	3,7 %
Scientifique	19,2 %	32,1 %	**35,9 %**	2,6 %	3,8 %	1,3 %	2,6 %
Économique	4,2 %	35,6 %	**45,2 %**	1,3 %	2,8 %	1,3 %	9,6 %
Journalistique	**40 %**	26,6 %	-	6,6 %	13,4 %	-	13,4 %
total	20,2 %	**39,5 %**	23,8 %	1,7 %	7,2 %	2,3 %	5,3 %

Table 16. Patrons syntaxiques des titres.

Il y a une majorité de syntagmes nominaux comme têtes des titres. Dans les corpus plus « techniques » il existe un nombre important de titres avec séparateurs initiaux. Dans le corpus journalistiques, il y a une majorité de phrases avec verbe conjugué.

3.7 Résumé

Nous avons vu, dans ce chapitre, une caractérisation de différents phénomènes qui, de par leurs caractéristiques (linguistiques, structurelles, etc.), sont susceptibles de poser des problèmes en analyse automatique. Au regard des évaluations menées sur quelques analyseurs existants (IFSP, SEXTANT, *cf.* chapitre 2) ces phénomènes ne sont pas pris en compte dans les grammaires de ces systèmes (ils ne le sont pas, non plus, pour d'autres analyseurs comme FDG ou FIPS).

Cependant, la présence non négligeable de ces phénomènes dans des corpus variés rend nécessaire leur prise en compte dans le cadre d'un analyseur robuste ayant pour objectif une analyse linguistique fine.

À la lumière de ces constats, le chapitre suivant présente une vue globale de notre approche. Ainsi, le modèle que nous proposons intègre, fondamentalement, une modélisation de l'ensemble des phénomènes étudiés jusqu'ici par le biais de différentes grammaires. L'analyseur devient alors un ensemble de grammaires modulaires et reconfigurables, dont l'application dépend des caractéristiques du texte en entrée.

Chapitre 4

Vue globale de l'approche

4.1 Introduction

D'après les constats établis jusqu'ici, il semble nécessaire de redéfinir les bases d'un système d'analyse syntaxique vraiment robuste, c'est-à-dire capable de traiter du texte tout venant quel que soit son domaine, avec beaucoup plus de finesse (précision des analyses) et d'homogénéité (résultats similaires pour les différents types de phrases).

Pour ce faire, nous avons eu à notre disposition un ensemble d'outils d'analyse linguistique automatique au sein du Centre de Recherche Européen de Xerox (XRCE). À l'heure actuelle, la plupart de ces outils se trouvent intégrés dans une plate-forme d'analyse nommée XIP (*Xerox Incremental Parser*) [AMCR02] dont les propriétés convienent parfaitement à nos objectifs : ouverture (les règles des grammaires sont facilement enrichissables), modularité (des grammaires peuvent être ajoutées ou enlevées), souplesse (il est possible d'articuler grammaires, lexiques et autres ressources).

Sur la base de XIP, nous avons créé notre modèle d'analyseur dans le but de faire face aux carences de la plupart des analyseurs existants, spécialement dans le traitement de phénomènes « complexes » au niveau linguistique et structurel (ponctuation, listes, titres, rattachement prépositionnel, etc.).

Par souci de clarté, nous appellerons dorénavant XIP la plate-forme d'analyse à la base de notre modèle et nous utiliserons XIP-F pour faire référence à l'analyseur du français de Xerox construit sur cette même plate-forme, avec ses propres grammaires.

De façon générale, notre modèle d'analyseur présente des différences importantes par rapport à l'analyseur XIP-F en ce qui concerne l'architecture du système et la conception des grammaires. Cependant, nous avons construit l'ensemble des grammaires qui réalisent l'analyse linguistique selon notre approche en prenant à la base –et en les raffinant– les grammaires pour les syntagmes noyau et les dépendances de base de XIP-F[1].

Dans ce contexte, l'objectif de ce chapitre est de décrire globalement notre travail ainsi que le formalisme à la base de XIP. Plus concrètement, la section 4.2 décrit les apports que nous avons introduits au niveau de la description et de la modélisation linguistique : ajout de pré-traitements spécialisés, stratégie d'analyse en deux étapes (modules noyau,

[1]Une liste exhaustive des syntagmes de base de XIP-F se trouve dans l'annexe B.2 ; pour les dépendances de base, voir l'annexe B.3.

modules spécialisés), développement complet des grammaires spécialisées et de la grammaire pour l'extraction du rattachement prépositionnel, lexicalisation de la grammaire de dépendance pour le rattachement prépositionnel par l'introduction d'une méthode d'apprentissage non-supervisé.

La description présentée dans les sections suivantes concerne fondamentalement le formalisme à la base de XIP ; nous faisons cette description par le biais de l'analyseur XIP-F existant[2]. Ainsi, la section 4.3 introduit les caractéristiques principales de la plateforme, en mettant l'accent sur l'architecture générale de XIP-F. La section 4.4 a pour objet la description détaillée de la représentation élémentaire des données dans le système.

La section 4.5 décrit le formalisme d'expression des différentes règles de la grammaire acceptées par XIP, spécialement pour les mécanismes concernant le découpage en syntagmes noyau de base et pour ceux liés à l'extraction de dépendances de base (qui ont servi de point de départ à notre travail). Le chapitre se conclut avec la section 4.6 qui récapitule et illustre, grâce à XIP-F, l'ensemble des étapes impliquées lors des différents niveaux de l'analyse linguistique.

4.2 Spécifications générales

Notre travail au niveau de la description et de la modélisation linguistique (et par là-même au niveau de l'architecture générale du système) intervient à deux niveaux :

- pré-traitements : identification de *phrases* (avant analyse morphologique) mettant en jeu des aspects typographiques, structuraux, etc. ;
- analyse syntaxique : création des grammaires pour le *chunking* et l'extraction de dépendances basée sur un découpage en différents modules linguistiques selon les phénomènes présents dans la phrase et lexicalisation des règles par apprentissage à partir d'exemples.

La figure 4.1 donne une vision générale de l'architecture de notre modèle. En italique, nous montrons les modules contenant des règles de grammaire existant dans l'analyseur XIP-F.

Comme le montre la figure, tout texte en entrée est tout d'abord soumis à des pré-traitements. En plus des traitements morphologiques habituels (segmentation, étiquetage, désambiguïsation) et à la différence du système XIP-F original, ces pré-traitements incluent une phase de repérage de phrases qui tient compte d'aspects liés à la visualisation de certaines parties du document (*cf.* 4.2.1).

Par la suite, le texte pré-traité est analysé par notre grammaire noyau et, dans des cas plus complexes, par des grammaires spécialisées que nous avons conçues suite à notre étude sur corpus présentée dans le chapitre précédent. Cette analyse « en deux étapes » est aussi une particularité de notre modèle. L'ensemble des grammaires réalise un découpage en segments (syntagmes noyau et autres structures particulières (*cf.* 4.2.2)) en tenant compte de différents phénomènes présents dans chaque phrase à traiter.

[2]Il s'agit du travail de S. Aït-Mokhtar, J. P. Chanod et C. Roux pour la définition du formalisme, et de C. Roux pour l'implémentation du moteur du système.

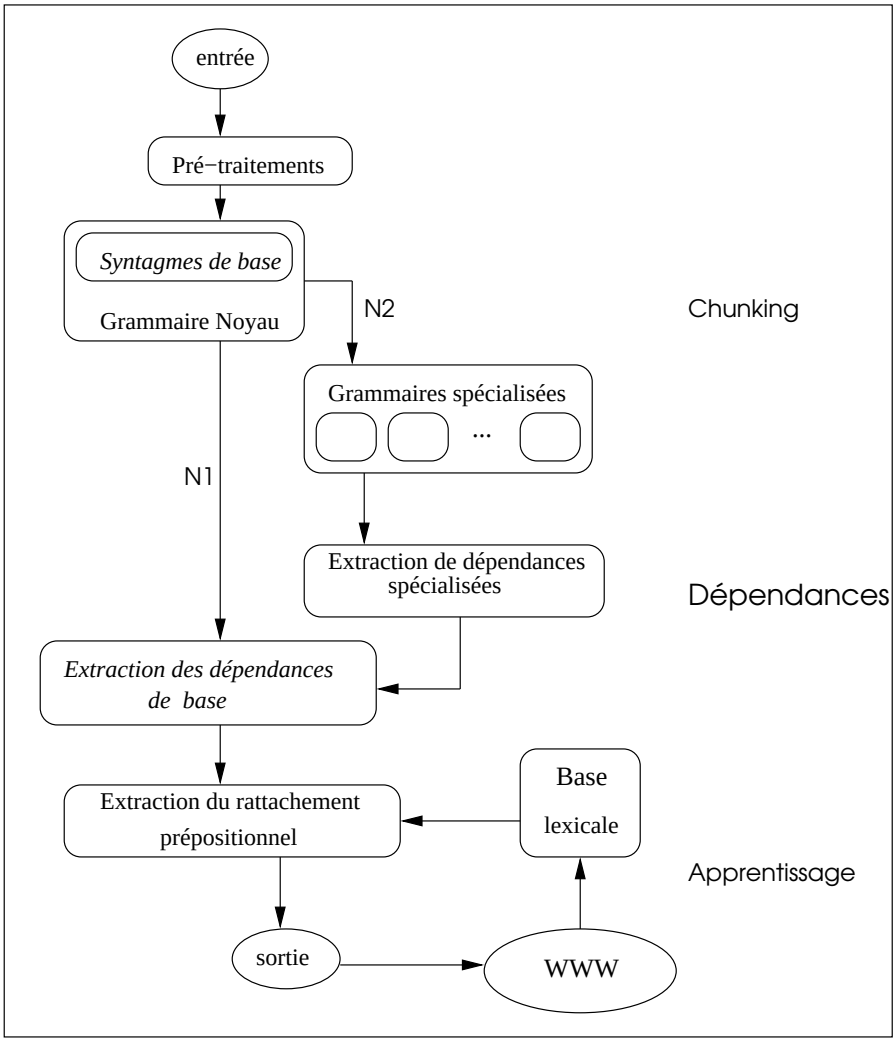

FIG. 4.1 – Architecture de notre modèle.

À la sortie de l'étape de marquage structurel, l'analyseur extrait des dépendances entre les têtes des différents segments repérés, dans le cas général. Cela constitue l'objectif principal de notre analyseur. Finalement, notre modèle comprend une étape d'apprentissage non supervisé qui utilise d'autres ressources comme le Web et des mesures statistiques (par exemple l'*information mutuelle*) dans le but d'améliorer une première sortie de l'analyseur lors de l'extraction de dépendances liées au rattachement prépositionnel (*cf.* 4.2.3).

Ainsi, nous nous sommes intéressée aux aspects suivants :

– l'**analyse pré-morphologique** : la préparation du texte en entrée en vue d'un meilleur découpage en unités de base,

– l'**analyse syntaxique** : la mise en place d'une méthodologie fondée, d'une part, sur la distinction de deux niveaux d'analyse et, d'autre part, sur l'acquisition automatique d'information lexicale permettant l'enrichissement d'une des grammaires de dépendance (celle pour l'extraction du rattachement prépositionnel).

Ces ajouts mettent l'accent sur la modularité et la reconfigurabilité des grammaires et rendent ainsi l'analyseur plus performant, en termes d'analyse linguistique, lors du traitement de textes hétérogènes tout venant (comme nous le montrons lors de l'évaluation).

Dans la suite de cette section, nous présentons brièvement ces aspects, qui feront l'objet des chapitres suivants : le début du chapitre 5 pour l'analyse pré-morphologique, la deuxième partie de ce même chapitre ainsi que la totalité du 6 pour l'analyse syntaxique à deux niveaux (grammaire noyau et grammaires spécialisées), et les chapitres 7 et 8 pour la lexicalisation des grammaires de dépendances.

4.2.1 Ajout de pré-traitements spécialisés

En tenant compte de certains problèmes de segmentation (*cf.* 4.6.1), nous avons jugé nécessaire de constituer un module d'identification et de marquage (balisage) de certains types de phrase. Ce module (qui précède la grammaire noyau, voir figure 4.1) constitue un *filtre* et s'avère un pré-traitement indispensable sur les corpus tout venant avant toute analyse linguistique proprement dite. Son objectif est l'identification et le marquage de deux sortes d'unités linguistiques : les listes et les titres. De par leurs caractéristiques, ces unités intègrent des notions ayant trait à la structure des documents et demandent une analyse syntaxique particulière. Une telle analyse ne peut être mise en œuvre dans le cadre de l'analyseur que si ces unités ont un marquage préalable.

Le module de pré-traitement prépare donc le corpus aux traitements syntaxiques ultérieurs[3]. Il reçoit en entrée un texte tout venant en format ASCII comme le suivant :

(1) *1.9.1-Le choix de la conception*

(...)

Ceci met en évidence la nécessaire prise en compte de l'activité globale et des interfaces de la situation réelle du travail, entre :

- la conception du matériel .
- la qualité de la peau à respecter .
- le poids des caisses d'emporte-pièce .
- la vitesse d'exécution (rendement) demandée .
- la conception du modèle à travailler .
- l'espace de travail et son encombrement, etc. .

[3]Du point de vue de l'implémentation, ce module est à l'heure actuelle un script `perl` d'environ 200 lignes pour le balisage de listes et de 30 pour les titres, *cf.* annexe C.

Ce phénomène d'interactions crée de nouvelles contraintes que les opératrices sont obligées de compenser par un engagement physique provoquant les douleurs qu'elles déclinent.

À la sortie du pré-traitement, le texte initial présente des marques spécifiques (balises) à des endroits particuliers. Ces balises sont traitées comme des ponctuations spéciales, mises au dictionnaire avec des propriétés adéquates et analysées par le système comme des unités lexicales portant une information indispensable à la bonne segmentation en *phrases*[4].

L'exemple suivant est le résultat du pré-traitement pour le texte de l'exemple précédent :

(2) `<tit>` *1* `<sp>.</sp>` *9* `<sp>.</sp>` *1 - Le choix de la conception*
`</tit>`
(...)
`<list>` *Ceci met en évidence la nécessaire prise en compte de l'activité globale et des interfaces de la situation réelle du travail, entre* `<sp> :</sp>`
- *la conception du matériel* `<sp>.</sp>`
- *la qualité de la peau à respecter* `<sp>.</sp>`
- *le poids des caisses emporte-pièce* `<sp>.</sp>`
- *la vitesse d'exécution (rendement) demandée* `<sp>.</sp>`
- *la conception du modèle à travailler* `<sp>.</sp>`
- *l'espace de travail et son encombrement, etc* `<sp>.</sp>`

`</list>`
Ce phénomène d'interactions crée de nouvelles contraintes que les opératrices sont obligées de compenser par un engagement physique provoquant les douleurs qu'elles déclinent.

En plus des balises `<list>` et `<tit>` (pour listes et titres, respectivement) il est nécessaire de masquer tout signe de ponctuation qui pourra être considéré ensuite par le parseur —plus tard— comme une fin de phrase (puisque le moteur de XIP que nous utilisons pour construire notre modèle d'analyseur traite les entrées selon un découpage « classique » par phrases). Ceci est fait par des balises `<sp>` (séparateurs).

L'ensemble des balises ajoutées lors de cette étape de pré-traitement sont considérées ultérieurement par le parseur comme des unités lexicales ; leurs étiquettes morphologiques sont prises en compte par les règles des grammaires spécialisées.

Il est à signaler que l'objectif de notre travail est d'analyser des corpus tout venant en format texte (ASCII). Nous n'utilisons pas, pour l'instant, des textes en format déjà enrichi avec des balises (SGML, HTML, XML, etc.). Les balises des corpus sous ces formats, bien qu'elles marquent les débuts et fins de listes et titres, répondent uniquement à des critères concernant la structure du texte[5].

[4] Cette stratégie est similaire à celle décrite par [BN81] dans le cadre d'un système réel de TA.

[5] Il est possible d'écrire un module de conversion du format HTML, en particulier pour les listes et les titres, en un format accepté par notre analyseur, mais nous avons laissé ce travail en dehors du cadre de la thèse.

Le marquage que nous proposons, en plus de ces critères, intègre des notions plus linguistiques, comme le marquage de la ponctuation forte, la prise en compte de séquences introductrices pour les listes, etc. Ces notions s'avèrent indispensables pour les traitements linguistiques ultérieurs.

4.2.2 Définition de deux niveaux d'analyse

L'analyseur produit comme résultat à partir d'un texte en entrée, un texte découpé en phrases (segmentées en syntagmes noyau) et une liste de relations de dépendance (entre les têtes des syntagmes des phrases).

Dans notre approche, ce résultat est produit grâce à l'application de différentes grammaires correspondant à des niveaux d'analyse différents. L'analyse par niveaux est une stratégie mise en œuvre dans certains formalismes existants, par exemple celui de J. Slocum [SBB+87]. Mais à la différence de cette approche et des approches d'autres analyseurs existants, ici ces niveaux d'analyse sont définis en fonction des caractéristiques structurelles de chaque phrase en entrée [GP01]. Nous définissons deux niveaux en fonction de la typologie de phrases que nous avons modélisée.

En effet, nous partons du principe, que nous vérifions en section 5.3, que dans tout corpus il existe des phrases présentant une structure syntaxique « simple » (représentée par la configuration `SVO` avec pas ou peu de modifieurs) et dont une analyse automatique précise (dans le sens anglais d'*accurate*) n'est pas difficile à obtenir. Nous appelons ce type de phrases des *phrases de premier niveau* (N1).

Exemples :

(3) *Les caches plastiques se trouvent sur le côté intérieur.*

(4) *Un appel en garantie paraît juridiquement impossible.*

Ce premier niveau d'analyse est modélisé par un ensemble de règles correspondant à une grammaire noyau.

Par ailleurs, les phrases qui présentent des phénomènes plus complexes à modéliser sont traitées par différents modules spécialisés. Ces phrases, que nous appelons des *phrases de deuxième niveau* (N2) présentent des caractéristiques liées à la présence de phénomènes comme la ponctuation et les espaces verticaux ou horizontaux ainsi qu'à la composition et organisation de leurs constituants (coordinations et rattachements prépositionnels ambigus, appositions, etc.).

(5) *L'Ecole des hautes études en sciences sociales (EHESS, 54, boulevard Raspail, Paris 6) accueille, du 28 janvier (17 h 30) au 6 mars, " Les images médiatiques et la ville ", une exposition comprenant 45 000 timbres de France, d'Allemagne, d'Espagne, d'Italie, de Grande-Bretagne et des pays de l'ex-bloc de l'Est.*

(6) *Sur les modèles jusqu'en 1965 voici les instructions à suivre :*
- Enlevez les enjoliveurs de roue.
- Soulevez la roue et tournez-la vers l'avant jusqu'à ce que le trou de réglage pratiqué dans le tambour se trouve en face d'une des deux molettes de réglages.

L'ensemble des modules ou grammaires spécialisées dans le traitement des phénomènes présents dans ces types de phrase correspond dans notre système au deuxième niveau d'analyse.

Premier niveau

Le premier niveau d'analyse a comme objectif une première analyse structurelle de chaque unité en entrée. Cette analyse permet tout d'abord le découpage en syntagmes noyau de base et par la suite le tri entre les phrases N1 et N2. Le tri se fait par un marquage de syntagmes noyau additionnels. Un ensemble d'indices sur l'ordre et la composition de ces syntagmes définissent formellement les types de phrase.

Par exemple, une structure `SVO` avec deux virgules au maximum est considérée N1, une structure `SVO` avec des parenthèses N2, une structure avec des espaces horizontaux N2, etc. (de plus amples détails sont donnés aux chapitres 5 et 6).

Les phrases identifiées comme N1 sont envoyées au module d'extraction de dépendances. En revanche, les phrases N2 sont dirigées vers des modules particuliers (deuxième niveau d'analyse) dans le but de prendre en compte avec précision les différents phénomènes linguistiques et structurels.

L'ensemble des règles constituant la grammaire noyau de notre modèle permet d'analyser environ 20 % des phrases des corpus sur lesquels nous avons effectué nos évaluations[6] avec un taux de précision qui se situe autour de 96 %. Nous développons en détail les caractéristiques de ces règles et les résultats obtenus grâce à cette grammaire dans le chapitre suivant de ce rapport.

Deuxième niveau

Le deuxième niveau d'analyse est constitué d'un ensemble de grammaires spécialisées dans le traitement de phénomènes complexes liés à certains domaines de spécialité ou à certains modes d'écriture. L'application de ces grammaires n'est effective que si le phénomène en question est présent dans l'unité en entrée (par exemple la grammaire des listes ne s'applique que si l'unité en entrée est ou contient une liste).

À ce niveau de l'analyse, les phrases arrivent partiellement analysées en syntagmes de base. Elles contiennent des structures plus complexes à modéliser du point de vue automatique et requièrent des analyses spécifiques de leurs structures (définition de nouveaux syntagmes noyau adaptés) et de leurs dépendances (création de nouvelles relations).

D'après une évaluation initiale de différents analyseurs (*cf.* 2.3) ainsi qu'une étude approfondie sur corpus (*cf.* 3), nous avons créé des grammaires spécialisées pour les traitements des phénomènes suivants :

- ponctuation,
- listes et énumérations,
- titres.

[6] Ces corpus sont décrits plus loin, *cf.* 5.3.4.

Deux constats nous ont amenée à étudier et à modéliser des phénomènes liés à la ponctuation dans le cadre de notre modèle d'analyseur. D'une part, les traitements spécifiques des marques de ponctuation en linguistique informatique sont peu nombreux (mis à part le travail de quelques auteurs [Bri94], [Whi95], [Jon94], [Jon96]). D'autre part, les phrases contenant des marques de ponctuation sont très fréquentes dans les corpus [GP00] et elles entraînent souvent des erreurs d'analyse.

De même, les caractéristiques structurelles des listes et énumérations ainsi que des titres sont souvent ignorées par les grammaires des systèmes existants, en général parce que des notions textuelles sont mises en cause (visualisation, structure du document, etc.). La non prise en compte de ces structures nuit à la précision linguistique de l'analyse fournie par l'analyseur ainsi qu'à sa robustesse.

La création de grammaires spécialisées dans le traitement de ces phénomènes a donc comme objectif de garantir la précision et la robustesse de l'analyseur. Nous développons en détail les caractéristiques et les résultats obtenus par l'ensemble de ces grammaires dans le chapitre 6.

4.2.3 Lexicalisation des grammaires de dépendances

Nous avons exposé plus haut que l'objectif principal de l'analyseur est l'extraction d'un ensemble de dépendances pour chaque unité de base identifiée dans le texte en entrée. Différentes heuristiques, encodées sous forme de règles, sont mises en œuvre pour extraire les éléments liés par une même dépendance dans le contexte d'une même phrase.

Les règles modélisant les dépendances intègrent des informations morpho-syntaxiques provenant de dictionnaires (par exemple, quelques traits liés à la rection) ou obtenues après l'analyse structurelle (par exemple, l'identification des « têtes » des syntagmes). Toutes ces informations sont cruciales pour l'identification d'une dépendance donnée.

Dans la plupart des cas, le système est déterministe : après évaluation de toutes les informations, une seule dépendance (celle considérée la plus probable) est extraite entre une « tête » et une unité qui en dépend.

Par exemple, dans la phrase suivante, les « têtes » des syntagmes nominaux *les échanges* et *les Etats Unis* sont potentiellement candidates à être dépendantes du verbe *dépasser* dans le contexte d'une dépendance de type sujet (définie comme une relation requérant un syntagme nominal devant d'un verbe fini dont il dépend, avec entre autres des contraintes d'accord en nombre).

(7) *En 1995, les échanges commerciaux entre l'Afrique du Sud et les Etats Unis ont dépassé les 738 millions de dollars.*

Après évaluation des informations syntaxiques, seul *échanges* s'avère le bon candidat (la présence de la conjonction de coordination élimine le deuxième candidat de cet exemple).

Toutefois, dans certains cas, l'extraction de dépendances se heurte à des ambiguïtés structurelles plus complexes : l'identification de la « tête » et/ou du dépendant n'est pas toujours possible avec des informations uniquement morpho-syntaxiques. C'est le cas des dépendances liées au rattachement prépositionnel, par exemple (*cf.* également l'exemple

donné en page 46 avec (20) et (21)) :

(8) *CL - Espan emploie 820 personnes dans 101 succursales.*

Dans cette phrase le syntagme prépositionnel *dans 101 succursales* doit être rattaché à une « tête » qui peut être ici le verbe du syntagme verbal *emploie* ou le nom du syntagme nominal *820 personnes*. Devant une telle structure, les analyseurs syntaxiques à base de règles existants, comme XIP-F, ne sont pas capables de produire le lien correct, car les informations dont ils disposent ne sont pas suffisantes pour établir le bon rattachement.

Dans le cas de XIP-F, la grammaire pour le rattachement prépositionnel est non déterministe : elle extrait tous les rattachements possibles[7]. Les dépendances extraites par cet analyseur pour la phrase précédente sont un modifieur verbal et un modifieur nominal :

```
VMOD(emploie,dans,succursales)
NMOD(personnes,dans,succursales)
```

Dans des cas comme celui-ci où la prise d'une décision n'est pas possible avec des informations uniquement morpho-syntaxiques, d'autres informations acquises par d'autres techniques s'avèrent nécessaires.

Ainsi, il existe des systèmes d'extraction de dépendances syntaxiques fondés sur des approches statistiques. L'extraction de dépendances se fait alors par des calculs probabilistes en utilisant des méthodes d'apprentissage à partir de corpus d'entraînement préalablement annotés. Toutefois, ces approches ne permettent pas non plus de déterminer les bons rattachements dans tous les cas (*cf.* chapitre 7, sous-section 7.2).

La prise en compte de ces constats nous a amenée à envisager un système « hybride » fondé sur l'utilisation de techniques d'apprentissage sur corpus à partir d'une première analyse produite par des grammaires à base de règles (à l'instar de [Sri97] et [BM97]).

L'idée principale est l'introduction d'information lexicale, sous forme de patrons de cooccurrence syntaxique, ainsi que des mesures statistiques (estimations de fréquences de rattachement) utilisées lors d'une étape supplémentaire d'analyse consacrée à la désambiguïsation des rattachements (voir figure 4.1).

Nous décrivons en détail cette méthode et les différents résultats des expériences mises en œuvre dans les chapitres 7 et 8.

Une vision globale de notre approche étant maintenant présentée, nous allons jusqu'à la fin de ce chapitre nous intéresser aux caractéristiques de la plate-forme XIP, avec des exemples produits par XIP-F, notre travail étant bati, à la base, sur cet analyseur.

4.3 Présentation de XIP

Le paradigme computationnel dans lequel s'inscrit le formalisme de XIP est celui de l'analyse syntaxique robuste à base de règles, dans la même orientation que d'autres

[7]Ce choix fait baisser la précision de l'analyse mais donne priorité au rappel.

systèmes comme PLNLP [Jen92] (approche continuée actuellement chez Microsoft), FDG [VJ96] (post-ENGCG [KVHA95]) ou IFSP [AMC97a].

Les différents processus linguistiques s'organisent de façon incrémentale : à partir d'un texte annoté morphologiquement, l'analyse syntaxique est fondée sur un découpage initial en syntagmes noyau (*chunking* [Abn96]) suivi d'une extraction de dépendances syntaxiques.

L'objectif final du système est de fournir une analyse en termes de dépendances qui pourront être utilisées par la suite dans d'autres applications comme la reconnaissance d'entités nommées, la désambiguïsation sémantique, l'apprentissage automatique de classes sémantiques, la résolution de la coréférence, la traduction automatique, etc.

Sur la base de XIP, il existe à l'heure actuelle deux analyseurs (français[8] et anglais) utilisés dans deux projets, l'un concernant l'extraction d'information dans le domaine de la biologie (projet BioTIP, précédemment Munnin [PRL00], [Pro01]), l'autre concernant l'enrichissement de documents à partir de données provenant d'une encyclopédie (projet CIRCE [JBR02]).

4.3.1 Architecture

XIP permet l'articulation de plusieurs modules pour l'analyse linguistique. Il accepte en entrée des données en format texte ou bien en format résultant d'un traitement préalable (il est ainsi *multi-input* [AMCR01]). Chaque module transforme et analyse le format donné en entrée (généralement il s'agit du texte tout venant). Dans le cas de XIP-F, les principales transformations concernent la désambiguïsation morphologique, la segmentation de séquences (en syntagmes noyau) et l'extraction de dépendances.

Si on considère XIP-F comme un analyseur en dépendances, il est alors composé de trois modules optionnels (prétraitement ou analyse morphologique, désambiguïsation à base de règles ou bien à base de HMM, module de *chunking*) et un module fondamental pour l'extraction de dépendances. Ainsi, il peut prendre un entrée un texte déjà segmenté en syntagmes noyau pour faire juste le calcul de dépendances (voir figure 4.2).

Ces quatre modules correspondent à :

- **module de normalisation, tokenisation et morphologie (NTM)** : il donne une forme typographiquement normalisée de l'entrée, segmente celle-ci en une séquence de « formes » (*tokens*, c'est-à-dire mots, symboles de ponctuation, balises, etc.) et assigne toutes les informations lexicales pour chaque forme identifiée (catégorie et traits morphosyntaxiques) ;
- **module de désambiguïsation contextuelle** : il désambiguïse les catégories des mots selon leur contexte d'apparition ;
- **module de segmentation** : il découpe les unités linguistiques principalement en séries de syntagmes noyau mais aussi en d'autres structures ;
- **module d'extraction de dépendances** : il identifie des liens syntaxiques entre les mots.

[8] Les performances sur machine Pentium III 1 GHz sont d'environ 2000 mots/sec. et de 10 Mo d'espace mémoire (grammaires uniquement, sans les lexiques).

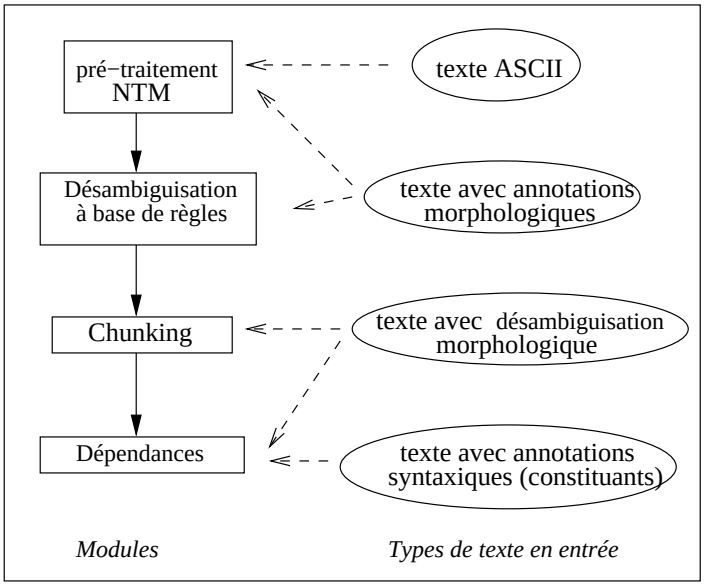

FIG. 4.2 – Architecture de l'analyseur XIP-F.

Les deux derniers modules sont ceux qui réalisent l'analyse syntaxique proprement dite.

La sortie de l'analyseur est un texte annoté, principalement avec des relations de dépendance entre les différents mots d'une *phrase* et optionnellement avec les marques de découpage en syntagmes noyau et les informations morphologiques plus ou moins riches. Différentes sorties sont données en exemple dans la section 4.6 et plus en détail dans l'annexe B.2.

4.4 Représentation élémentaire des données analysées

L'analyse syntaxique finale du parseur (ensemble de relations de dépendance) est obtenue à partir des différents processus linguistiques effectués dans les différents modules. La représentation des données à chaque étape se fait sous la forme d'un arbre syntaxique partiel.

4.4.1 Arbre syntaxique

La représentation de base des données à tous les niveaux de l'analyse est un nœud faisant partie d'un arbre syntaxique. On distingue deux types de nœuds : les nœuds lexicaux (*mots*) et les nœuds non lexicaux (catégories, paires attribut-valeur, etc.). Les nœuds lexicaux dominent immédiatement les unités lexicales, ces unités étant les feuilles de l'arbre. Les nœuds non lexicaux dominent les nœuds lexicaux. Nous allons préciser ces notions dans la suite de cette section.

Le texte en entrée est subdivisé en unités (ces unités représentant les *phrases*). Chaque unité à traiter est alors considérée comme une séquence de nœuds terminaux (s'il n'y a

pas de structures en constituants) ou comme une séquence de nœuds de constituants (séquences de sous-arbres).

Les limites entre les unités sont définies par des séquences de nœuds du texte en entrée (voir sous-section consacrée au pré-traitement).

Pour XIP-F, les nœuds lexicaux sont spécifiés à partir des informations fournies par l'analyse morphologique et les processus de désambiguïsation des unités lexicales. À partir de l'analyse morphologique, l'analyseur construit un premier arbre syntaxique : chaque unité lexicale a pour parent un nœud dont l'étiquette est celle de la partie du discours associée à l'unité lexicale.

Par exemple pour les trois premiers nœuds lexicaux de la phrase (9), l'analyse morphologique (module NTM) donne le résultat reproduit par la suite :

(9) *"Les variations soumettent les particules à des mouvements vibratoires ."*

```
Les     le      +Acc+InvGen+PL+P3+PC
Les     le      +InvGen+PL+Def+Det+DET_PL

variations      variation       +deSN+dansSN+Fem+PL+Noun+NOUN_PL

soumettent      soumettre       +se+aSN+SN+avoir+SubjP+PL+P3+Verb+VERB_P3PL
soumettent      soumettre       +se+aSN+SN+avoir+IndP+PL+P3+Verb+VERB_P3PL
```

La désambiguïsation morphologique par XIP permet la construction d'un arbre syntaxique qui, à ce stade, a la forme représentée par la figure 4.3.

```
 |        |        |
DET     NOUN     VERB
 +        +        +
 |        |        |
Les variations soumettent
```

FIG. 4.3 – Arbre syntaxique de nœuds lexicaux.

Pour XIP-F, les étiquettes possibles pour les nœuds lexicaux correspondent aux catégories morphologiques classiques, principalement :

- `NOUN` (nom),
- `DET` (déterminant),
- `PRON` (pronom),
- `ADJ` (adjectif),
- `VERB` (forme verbale, participes inclus),
- `ADV` (adverbe),
- `PREP` (préposition),

- `COORD` (conjonctions de coordination),
- `CONJ` (autres conjonctions),
- `NUM` (numéraux cardinaux),
- `PUNCT` (ponctuation ne marquant pas une fin de phrase),
- `SENT` (ponctuation marquant une fin de phrase).

À partir de cette représentation, les différentes règles de grammaire intégrant le module de *chunking* permettent de créer un arbre de *chunks* : l'arbre syntaxique devient à ce stade un arbre de syntagmes noyau.

Chaque syntagme noyau est un nœud non lexical[9]. Par exemple, un syntagme nominal noyau (`NP`) est un syntagme qui a comme limite droite un nom ou pronom. Une proposition finie noyau (*sentence clause* dans XIP-F, marquée SC) est la partie d'une proposition ayant pour noyau un verbe conjugué.

Le nœud maximal de chaque arbre (et donc de chaque *phrase*) est le nœud virtuel `GROUPE` (finissant par une marque de fin de phrase).

Pour l'exemple précédent, la représentation syntaxique à ce stade est l'arbre donné dans la figure 4.4 :

```
          SC
    +--------------+
    |              |
    NP            FV
 +-------+         +
 |       |         |
DET    NOUN       VERB
 +       +         +
 |       |         |
Les variations soumettent
```

FIG. 4.4 – Arbre syntaxique de nœuds non lexicaux.

Ce même exemple se représenterait comme l'arbre de la figure 4.5 s'il y avait un deuxième `NP` et une ponctuation forte finale (*"Les variations soumettent les particules."*).

4.4.2 Système de traits

Aux nœuds de l'arbre syntaxique, qu'ils soient lexicaux ou non, sont associés des « traits ». Un trait vise à exprimer une propriété associée au nœud concerné. Il s'agit principalement d'une information d'un des quatre types suivants :

- typographique : majuscules `[maj :+]`, etc. ;
- morphologique : genre `[fem :+]`, nombre `[plu :+]`, personne `[p3 :+]`, etc. ;

[9] Pour rappel (*cf.* chapitre 3, sous-section 3.2 *Précisions terminologiques*), un syntagme noyau ou *chunk* est la partie d'un syntagme qui va de son début jusqu'à son noyau inclus. La notion de noyau correspond à la notion de « tête ».

```
                        GROUPE
           +-------------------+------------+
           |                   |            |
           SC                  NP          SENT
    +--------------+        +-------+       +
    |              |        |       |       |
    NP            FV       DET     NOUN     .
 +-------+         +        +       +
 |       |         |        |       |
DET    NOUN       VERB     les   particules
 +       +         +
 |       |         |
Les variations soumettent
```

FIG. 4.5 – Arbre syntaxique de nœuds non lexicaux avec nœud maximal.

- syntaxique : premier élément d'une phrase `[start :+]` ou d'un syntagme `[first :+]`, rection (par exemple pour la préposition *a*) `[sfa :+]`, etc. ;
- sémantique : temps `[time :+]`, etc. .

Le formalisme de XIP est comme celui de certains langages de programmation où chaque trait (ou catégorie) doit être déclaré. Le système de traits permet de spécifier des catégories, une catégorie étant un ensemble de nœuds définis par les traits qui doivent ou non être associés aux éléments de cet ensemble.

De plus, les traits de la tête d'un syntagme noyau sont souvent hérités par le nœud du syntagme. C'est-à-dire que les traits caractérisant une unité lexicale tête d'un syntagme caractérisent aussi le syntagme en entier. Par exemple, le trait `[noun :+]` de *variations* est hérité par le nœud du `NP` qu'il forme avec le déterminant.

Pour la phrase en exemple, les traits associés aux premiers nœuds sont, entre autres, ceux de la figure 4.6 (selon XIP-F, car d'autres traits peuvent être ajoutés ou enlevés selon le modèle d'analyseur) :

```
NP              [maj:+,noun:+,det:+,start:+,first:+]
 DET            [maj:+,masc:+,fem:+,pl:+,def:+,det:+,start:+,first:+]
   Les          [masc:+,fem:+,pl:+,def:+,det:+,start:+,first:+]
 NOUN           [fem:+,pl:+,p3:+,noun:+,last:+]
   variations   [fem:+,pl:+,p3:+,noun:+,last:+]

FV              [fin:+,verb:+]
 VERB           [se:+,sfa:+,pl:+,p3:+,pre:+,subj:+,ind:+,verb:+,first:+]
  soumettent    [se:+,sfa:+,pl:+,p3:+,pre:+,subj:+,ind:+,verb:+,first:+]
```

FIG. 4.6 – Ensemble de traits associés.

Les informations associées aux différents traits sont déclarées dans la grammaire sous la forme suivante :

```
attribut1:{valeur1, valeur2,..., valeurN}
```

La forme `attribut1` est le nom de l'attribut et `valeur1, valeur2,...,valeurN` l'ensemble de valeurs différentes que peut prendre l'attribut donné[10].

L'affectation de traits se fait, au niveau des nœuds lexicaux, par la traduction des étiquettes fournies par l'analyseur morphologique. Par exemple, pour le déterminant *Les*, l'analyseur morphologique donne le résultat suivant, la première colonne étant la forme de surface, la deuxième le lemme et la troisième l'ensemble d'informations morphologiques :

```
Les     le      +InvGen+PL+Def+Det+DET_PL
```

Cette analyse donne lieu, dans le système XIP, à la création d'un nœud lexical dominant l'unité *Les* qui aura l'étiquette `DET` et les traits suivants associés :

```
DET     [masc:+,fem:+,pl:+,def:+,det:+]
```

Ces traits sont « lexicaux », mais d'autres traits « syntaxiques » s'ajouteront à ceux-ci pendant l'analyse syntaxique. En effet, un autre mode d'assignation de traits est fourni par les règles de XIP elles-mêmes. Il est possible, dans une règle XIP, en même temps qu'on fait référence à un nœud, d'affecter un trait à ce nœud (`[attribut=valeur]`) ou de supprimer un trait déjà affecté (`[attribut= ~]`).

4.4.3 Unités de base

À partir d'un texte annoté morphologiquement, le formalisme XIP permet d'identifier dans un premier temps les *unités de base* : il segmente le texte en entrée en unités à traiter. L'identification des limites de chaque unité permettra de créer un ensemble de séquences de nœuds.

Dans le cas de XIP-F, des *phrases* sont identifiées selon la notion classique[11], c'est-à-dire des propositions finissant par une ponctuation forte. Les fins de phrase correspondent alors aux marques de ponctuation suivantes : point [.], point d'interrogation [?], point d'exclamation [!], deux points [:] ou point virgule [;] suivis de retour à la ligne. Deux espaces verticaux (deux retours à la ligne) sont aussi considérés comme fin de phrase.

Pour tout découpage en unités de base, le texte en entrée (avec des informations morphologiques) est traité séquentiellement de droite à gauche (définition d'un point de segmentation puis de son contexte gauche). Un ensemble de règles de grammaire permet de définir les points de segmentation ainsi que les différents contextes possibles.

L'exemple suivant décrit un type d'unité de base dans XIP-F :

[10] Pour une description détaillée du système de traits de XIP-F, voir [Tro01].

[11] XIP permet de définir d'autres unités de base que la phrase au sens classique, *cf.* section 5.2 et aussi [Tro01].

```
sent[form:fpfin]
```

Cette règle définit une unité de base comme une séquence de nœuds finissant par un nœud `sent` ayant le trait `fpfin` (point final).

Voici un exemple plus complexe que nous avons défini dans notre modèle d'analyseur :

```
?+[form:~foquotes,form:~fcquotes,form:~fcpar,form:~fopar],
 verb,
 ?+[form:~foquotes,form:~fcquotes,form:~fcpar,form:~fopar],
 sent[form:fpfin]
```

Selon cette règle, une unité de base est une séquence d'au moins un nœud qui ne contient pas les traits `foquotes`, `fcquotes`, `fcpar`, `fopar` (guillemets ouvrants et fermants et parenthèse ouvrante et fermante, respectivement). Cette séquence doit être suivie d'un nœud `verb`, suivi d'un nœud `sent` avec le trait (`fpfin`) avec entre les deux les mêmes contraintes que pour la séquence initiale.

4.5 Formalisme d'expression des règles de la grammaire

L'analyseur est constitué d'un ensemble de règles de grammaire modélisant des structures et des fonctions syntaxiques potentiellement présentes dans une phrase. En ce qui concerne leur écriture et application, ces règles sont organisées par couches : une règle écrite au début de la grammaire s'appliquera aussi en premier lieu. L'ordre d'écriture est pertinent. De plus, les règles ne sont pas récursives : une certaine récursivité est induite avec les différents niveaux de règles.

L'application d'une règle est définitive : il n'y a pas de retour en arrière (*backtracking*) ; les syntagmes résultants ne sont pas supprimés et sont passés tels quels aux niveaux suivants.

Les règles de la grammaire utilisent les opérateurs décrits dans la table suivante :

Opérateur	*Symbole(s)*	*Exemple*
Disjonction	;	`ADJ ; ADV`
Concaténation	,	`NP , AP`
Optionnalité	( ) * +	`(ADV), ADJ*, NOUN+`
Catégorie quelconque ou rien	?	`NP,?*, FV`
Exploration d'un sous-arbre	{ }	`NP{?*, NOUN}`

Table 1. Opérateurs de base des règles.

Le premier exemple définit un nœud avec un adjectif *ou* un adverbe ; le deuxième un nœud avec un `NP` *suivi d'*un `AP`. Le troisième exemple dénote *un* ou *zéro* adverbe, *zéro* ou *plusieurs* adjectifs, *un* ou *plusieurs* noms (opérateur étoile de Kleene). Le quatrième exemple décrit un nœud constitué d'un `NP`, suivi de *zéro* ou *plusieurs catégories quelconques*, suivi d'un `FV`, et finalement le dernier exemple permet l'accès d'un nœud *à l'intérieur d'un sous-arbre* (le nom à l'intérieur d'un `NP`).

Il existe différents types de règles pour les différents traitements linguistiques. Nous présentons par la suite les types de règles pour les deux tâches principales de l'analyseur : le découpage en syntagmes noyau (*chunking*) et l'extraction de dépendances.

4.5.1 Règles de découpage en syntagmes noyau

Pour la description du découpage en syntagmes noyau, il existe plusieurs types de règles, chacun d'eux étant adapté à une tâche précise[12] :

- les règles de Dominance Immédiate,
- les règles de Précédence Linéaire,
- les règles de Séquence.

Les règles de Dominance Immédiate (DI), enrichies avec des traits, sont utilisées pour le marquage de segments ; les règles de Précédence Linéaire (PL) permettent de gérer l'ordre des éléments d'une unité et enfin les règles de Séquence sont utilisées pour définir une série ordonnée d'unités.

Règles de Dominance Immédiate

Les règles de DI définissent des ensembles de nœuds d'ordre libre. Elles ont la forme suivante :

```
n> X -> |C| Y, Z |C'|;
```

Le numéro de couche est indiqué par `n>`, `X` représente le nœud défini et `Y, Z` les éléments le définissant. `|C|` et `|C'|` représentent les contextes droit et gauche, respectivement. Ce sont des contextes optionnels.

Dans ce type de règle, la partie droite n'est pas ordonnée, mais des contraintes de position peuvent être ajoutées grâce aux traits `first` et `last`, par exemple :

```
2>NP -> Det[first:+], Adv*, AP, Noun[last:+];
```

Cette règle permet de définir un syntagme nominal (`NP`) qui contient un déterminant comme limite gauche, peut éventuellement avoir des adverbes, contient un syntagme adjectif noyau (`AP`) et doit avoir comme limite droite un nom. Sans ces traits `first` et `last`, elle autoriserait tout type de combinaison entre les éléments du syntagme (`Adv Det Noun`, `Noun Det Adv Adv AP Adv`, etc.).

Les règles sont écrites et organisées par couches et s'appliquent selon l'ordre établi par ces couches. Dans le cas précédent, le `NP` est défini lors de la deuxième couche (après les syntagmes adjectivaux `AP`). Si deux règles sont concurrentes, c'est la première écrite qui s'applique.

[12] On remarque des similitudes avec la typologie de règles du formalisme GPSG [GKPS85] qui est, en partie, à l'origine du système [Rou96]. Il y a aussi des similitudes avec l'analyseur IFSP [AMC97a] (les règles ne sont pas récursives, elles sont organisées en séquences, si une règle échoue l'analyse continue, etc.).

Des exemples de ces règles se trouvent dans la première partie de l'annexe D.1 (règles de la grammaire noyau correspondant à la définition des segments postérieur et antérieur au `NP` sujet et pour le segment postérieur au `FV` principal) et dans D.2 pour la définition des énumérations à trois items.

Règles de Précédence Linéaire

Les règles PL sont toujours utilisées en conjonction avec les règles DI. Elles permettent de définir un certain ordre dans les éléments d'un ensemble désordonné et se présentent sous la forme :

```
n> X < Y
```

`n>` indique le numéro de couche, `X` est l'élément antécédant `Y`. Par exemple :

```
2> [det:+] < [num:+]
```

Cette règle PL en concret implique qu'un déterminant `det` doit apparaître devant un numéral `num`.

Pour l'écriture de nos grammaires, nous avons utilisé très peu de règles de ce type car les contraintes qu'elles expriment peuvent souvent être exprimées avec des traits.

Règles de Séquence

Finalement, les règles de Séquence décrivent des ensembles de nœuds ordonnés. Elles ont la même syntaxe que les régles DI avec une différence quant à la forme (l'élément instancié est séparé des éléments le définissant par le symbole =) et quant à la sémantique (il n'est pas nécessaire de déclarer chacune des catégories apparaissant dans un segment). Il est aussi possible de définir des contextes droit et gauche optionnels.

```
n> X = |C| Y, Z |C'|;
```

Ces règles mettent en œuvre le principe des règles hors-contexte et sont donc beaucoup plus souples. Par exemple, la règle suivante décrit un nœud `PR` (parenthèse)[13].

```
10>PR = punct[form:fopar], ?*, punct[form:fcpar].
```

Ce type de syntagme est formé par une marque de ponctuation initiale qui a le trait `fopar` et une marque de ponctuation finale avec le trait `fcpar`. Entre ces deux marques, il peut y avoir rien ou tout autre symbole plusieurs fois. L'ordre des éléments, tout au moins le premier et le dernier, est ici fixé.

De nombreux exemples de règles de Séquence se trouvent dans l'annexe D, pour les phrases de niveau 1 dans D.1 et pour la plupart de syntagmes spécialisés dans D.2.

[13] Ce type de nœud n'existe pas dans l'analyseur XIP-F pré-existant : il fait partie de nos apports (voir 4.2).

4.5.2 Règles d'extraction de dépendances syntaxiques

Une dépendance est une relation n-aire qui connecte des nœuds selon une relation prédéfinie (relations syntaxiques standard —sujet, objet— ou des relations plus complexes).

Tout comme les règles de dominance immédiate et les règles de séquence pour le *chunking*, ces règles s'appliquent séquentiellement de façon incrémentale (le nombre de la couche n'est pas explicite pour ce type de règles).

Les règles d'extraction de dépendance ont la forme suivante :

```
if (C)
X(#1,#2) = Y, Z.
```

Le mot clé `if` introduit une condition (optionnelle) sur l'entrée susceptible d'instancier les règles. Il peut s'agir des conditions sur l'existence ou la non-existence d'une relation, sur la présence ou absence d'un trait, etc. Plusieurs conditions peuvent être exprimées ; dans ce cas, elles sont regroupées dans des formules complexes à l'aide des opérateurs booléens de conjonction (&), disjonction (|) ou négation (~). Par exemple :

```
if (~A(#1,#3,#4) & ~A[mf1:+](?,?,#4))
A[mf2=+]((#1,#3,#4)) =
FV{?*,#1[last:+]}, (adv), PP{?*,#3[prep:+,sfde:~], ?*, NP{?*,#4[last:+]} }.
```

La condition de la règle en exemple implique l'absence d'une relation `A` avec les arguments `#1,#3,#4` et l'absence d'une relation `A` ayant le trait `[mf1 :+]` et le troisième argument `#4`. Les deux conditions doivent être vérifiées pour que la règle s'applique.

Dans le corps de la règle, l'expression `X(#1,#2)` représente la dépendance à créer et `Y, Z.` les éléments la définissant. Il peut s'agir de nœuds lexicaux (par exemple, `adv`), de nœuds non-lexicaux (`NP`) ou d'opérateurs (`?*`). L'utilisation de traits associés permet de contraindre l'extraction des dépendances.

Pour l'exemple précédent, la dépendance `A(#1,#3,#4)` avec le trait `mf2` est créée avec les arguments suivants : un verbe fini (dernier élément d'un syntagme verbal `FV`) et la préposition (autre que *de*) et le nom d'un syntagme prépositionnel `PP`. Entre la préposition et le nom du `PP` il peut y avoir zéro ou plusieurs éléments quelconques.

Des exemples de ce type de règles pour l'extraction de dépendances se trouvent dans l'annexe D, concrètement dans D.3 (règles des grammaires de dépendances spécialisées) et D.6 (règles pour les dépendances liées au rattachement prépositionnel).

4.6 Traitements linguistiques de XIP-F

Les différentes étapes de l'analyse transforment le texte en entrée : selon les informations exprimées dans les règles, des annotations spécifiques sont introduites.

4.6.1 Segmentation

La première transformation réalisée par l'analyseur est le découpage du texte d'entrée en unités de base[14].

Le texte qui suit est découpé par XIP-F en huit unités de base (figure 4.7). Nous avons marqué par `[V]` les espaces verticaux ou retours à la ligne dans le texte original et par `<n>` le comptage d'unités à la sortie du segmenteur.

(10) *I. Dispositions concernant la qualité* `[V] [V]`

Article. 3 - Dans toutes les catégories , compte tenu des dispositions particulières prévues pour chaque catégorie et des tolérances admises , les bulbes doivent être entiers, sains (sont exclus les produits atteints de pourriture ou altérations), propres et pratiquement exempts de matière étrangère visible. `[V]`

Les échalotes doivent présenter un développement et un état tels qu' ils leur permettent : `[V]`

- de supporter un transport et une manutention ; `[V]`

- d' arriver dans des conditions satisfaisantes au lieu de destination. `[V]`

Article . 4 - Les échalotes font l' objet d' une classification en deux catégories définies ci-après . `[V]`

```
<1> I. Dispositions concernant la qualité
<2> Article.
<3> 3 - Dans toutes les catégories , compte tenu des dispositions
    particulières prévues pour chaque catégorie et des tolérances admises
    , les bulbes doivent être entiers, sains ( sont exclus les produits
    atteints de pourriture ou altérations ), propres et pratiquement
    exempts de matière étrangère visible.
<4> Les échalotes doivent présenter un développement et un état tels qu'
    ils leur permettent :
<5> - de supporter un transport et une manutention ;
<6> - d' arriver dans des conditions satisfaisantes au lieu de
    destination.
<7> Article .
<8> 4 - Les échalotes font l' objet d' une classification en deux
    catégories définies ci-après .
```

FIG. 4.7 – Sortie du segmenteur en unités de base (*phrases*).

[14]Pour rappel, le texte contient déjà à cet stade les informations provenant de l'analyse morphologique, c'est-à-dire les étiquettes des parties du discours (*POS tags*) enrichies d'autres informations (genre, nombre, etc.).

Il est facile de constater des problèmes de découpage étant donné, d'une part, l'ambiguïté de quelques marques de ponctuation (par exemple le point [.] ou les deux points [:]) et, d'autre part, la non prise en compte d'aspects liés à la visualisation de certaines parties du texte (par exemple avec les items des listes). En effet, il devrait y avoir, selon nous, un découpage en quatre unités (nous traitons en détail ces problèmes plus loin, *cf.* 5.2).

4.6.2 Désambiguïsation morphologique

Une fois le texte découpé en unités, les règles de sélection contextuelle (désambiguïsation morphologique) s'appliquent sur chaque séquence. Ces règles permettent la levée des ambiguïtés dans des cas où plusieurs étiquettes sont associées à une même unité lexicale.

Par exemple, pour la huitième phrase du texte identifiée par l'analyseur dans le texte en exemple, pour le nœud *une* (du syntagme *d'une classification*), le module morphologique NTM fournit les étiquettes morphologiques suivantes :

```
une     un     +Masc+SG+Card+Noun+NUM
une     un     +Masc+SG+Indef+Pro+PRON
une     un     +Masc+SG+Indef+Det+DET_SG
```

Dans ce cas, la levée de l'ambiguïté permet de ne garder que la dernière occurrence (DET). Par la suite, à chaque nœud lexical est associé une liste d'attributs, extraits de l'étiquette morphologique et/ou ajoutés par des règles.

La table 2 montre un exemple de la sortie à ce stade (nous avons simplifié et structuré les informations pour une meilleure lisibilité) :

Lemme	*Catégorie*	*Traits associés*
4	NUM	[start :+,first :+]
-	PUNCT	[form :fhyph]
le	DET	[maj :+,masc :+,fem :+,pl :+,def :+]
échalote	NOUN	[fem :+,pl :+]
faire	VERB	[form :ffaire,pl :+,p3 :+,pre :+,ind :+,verb :+]
le	DET	[masc :+,fem :+,sg :+,def :+]
objet	NOUN	[masc :+,sg :+,p3 :+]
de	PREP	[form :fde,prep :+]
un	DET	[masc :+,sg :+,indef :+,det :+]
classification	NOUN	[fem :+,sg :+]
en	PREP	[form :fen,prep :+]
deux	NUM	[card :+,num :+]
catégorie	NOUN	[fem :+,pl :+]
défini	ADJ	[fem :+,pl :+]
ci-après	ADV	[adv :+]
.	SENT	[last :+]

Table 2. Informations après la désambiguïsation morphologique.

Durant l'étape de désambiguïsation morphologique, des règles appartenant à des grammaires locales peuvent aussi s'appliquer ; par exemple des grammaires d'identification d'entités nommées, de marquage de dates, etc.

4.6.3 Découpage en syntagmes noyau

À partir de toutes les informations morphologiques obtenues jusqu'ici, les différentes règles de *chunking* s'appliquent de façon incrémentale pour construire un arbre de syntagmes noyau. L'ordre de marquage des syntagmes est le suivant : `AP, NP, PP`, suivis des syntagmes verbaux `FV, IV, GV`, suivis des débuts de proposition `BG` et enfin les propositions `SC`.

La sortie standard fournie par XIP-F est parenthésée avec les lemmes (mais le formalisme permet d'obtenir d'autres options, contenant plus ou moins d'information morpho-syntaxique[15]).

Pour la phrase en exemple, voici la sortie standard obtenue (chaque clef `{ }` entoure un syntagme noyau) :

```
8>GROUPE{SC{NP{4} - NP{Les échalotes} FV{font}} NP{l' objet}
PP{d' NP{une classification}} PP{en NP{deux catégories}}
AP{définies} ci-après .}
```

FIG. 4.8 – Sortie du segmenteur en syntagmes noyau (*chunks*).

On remarque des syntagmes imbriqués, notamment pour les syntagmes prépositionnels (`PP`), les propositions (`SC`) et le groupe maximal prédéfini (`GROUPE`).

4.6.4 Extraction de dépendances

Finalement, plusieurs dépendances sont extraites en utilisant les informations morpho-syntaxiques obtenues jusqu'ici[16].

Les dépendances sont des relations majoritairement binaires (à l'exception des relations concernant le rattachement prépositionnel). Le premier élément est toujours la « tête » et le deuxième l'élément dépendant.

L'extraction de dépendances s'effectue aussi de façon incrémentale. Les dépendances les plus sûres sont extraites en premier (`subj` et arguments régis par un verbe, un nom ou un adjectif) ; elles sont suivies des modifieurs et d'autres dépendances (éléments coordonnés, coréférencés, déterminés, etc.).

L'ensemble des dépendances principales produites pour la phrase en exemple est donné par la figure 4.9.

[15]Voir l'annexe B.2 pour différents types de sorties possibles.

[16]Voir l'annexe B.3 pour une description détaillée de l'ensemble de dépendances potentiellement fournies par XIP-F et leurs traits associés.

```
SUBJ_NOUN(font,échalotes)
VARG_NOUN_DIR(font,objet)
VMOD_ADV(font,ci-après)
NMOD_RIGHT_ADJ(catégories,définies)
NMOD_NOUN_INDIR(classification,en,catégories)
NMOD_NOUN_INDIR(objet,en,catégories)
NARG_NOUN_INDIR(objet,d',classification)
PREPOBJ_CLOSED(d',classification)
PREPOBJ_CLOSED(en,catégories)
DETERM_DEF_NOUN_DET(Les,échalotes)
DETERM_DEF_NOUN_DET(l',objet)
DETERM_NOUN_DET(un,classification)
DETERM_NUM_NOUN(deux,catégories)
STRAYNP_LEFT(font,4)

4 - Les échalotes font l' objet d' une classification
en deux catégories définies ci-après .
```

FIG. 4.9 – Sortie de l'extracteur de dépendances.

Dans l'exemple les relations extraites sont, entre autres : `SUBJ` (sujet) entre le verbe *font* et le nom *échalotes*, la relation `VARG` (argument verbal via nom) entre *font* et *objet*, et `VMOD` (modifieur verbal via adverbe) entre *font* et *ci-après*, etc.

4.7 Résumé

En partant des constats effectués dans les premiers chapitres, nous avons donné dans le début de celui-ci une vision globale de notre modèle d'analyseur robuste qui a pour objectif le traitement de texte libre de différents domaines avec des résultats précis et homogènes. Nous avons décrit brièvement les points plus significatifs de notre modèle, à savoir, une reconfiguration automatique de différents modules grammaticaux selon le type de phénomène à traiter et une réutilisation des résultats obtenus lors d'une première analyse pour améliorer les sorties initiales des grammaires de dépendances en ce qui concerne le rattachement prépositionnel.

Dans un deuxième temps, dans ce chapitre, nous avons aussi décrit les caractéristiques de la plate-forme XIP, par le biais de l'analyseur XIP-F pré-existant, notre modèle d'analyseur étant bâti sur cette plate-forme. Nous avons tout d'abord décrit l'architecture générale ainsi que le formalisme à la base de la représentation des données et de la définition des règles de grammaire. Nous avons également montré les traitements effectués par XIP-F dans les différentes étapes de l'analyse linguistique (segmentation, désambiguïsation morphologique, découpage en syntagmes noyau, extraction de dépendances) ainsi que les sorties obtenues à chaque stade.

Dans ce contexte, les parties suivantes de ce rapport présentent en détail la définition et l'implémentation du modèle d'analyseur que nous proposons sur la base de XIP.

Deuxième partie

Spécialisation et reconfigurabilité des grammaires

Chapitre 5

Pré-traitements et premier niveau d'analyse

5.1 Introduction

Après une présentation générale de la plate-forme à la base de notre modèle et des apports proposés dans le cadre de notre travail, ce chapitre s'articule autour de deux points fondamentaux.

D'une part, la section 5.2 présente notre module de pré-traitement qui a comme objectif l'identification formelle de deux types d'unités linguistiques dans des textes tout-venant (en format ASCII). La méthode mise en œuvre pour le repérage et marquage de ces unités est décrite en détail. À la lumière des résultats obtenus, ce module s'avère indispensable pour les traitements linguistiques postérieurs.

D'autre part, la section 5.3 présente notre grammaire noyau, premier module d'analyse syntaxique proposé dans notre approche. Nous justifions la nécessité d'un tel module dans le cadre de l'ensemble de l'analyseur et nous décrivons les caractéristiques des phrases modélisées par cette grammaire. Finalement, nous présentons ses principales étapes en termes de méthodologie d'analyse pour le marquage structurel et pour l'extraction de dépendances.

5.2 Pré-traitements

Nous définissons comme *pré-traitements* l'ensemble des opérations formelles réalisées sur le texte en entrée avant toute analyse purement linguistique. En effet, nous montrons (5.2.1) que des informations sur la structure du document sont nécessaires pour une bonne analyse linguistique. Ces informations ne peuvent être obtenues que par une étape préalable d'identification et de marquage de certaines unités dans le texte [Sri93].

5.2.1 Justifications

Traditionnellement, la *phrase* est considérée comme l'unité de base de la grammaire et par conséquent de l'analyse syntaxique. Elle est souvent définie dans des termes suivants :

"Unité de communication linguistique (...) suivie d'une pause importante. Dans le langage écrit, cette pause importante est généralement représentée par un point. " [Gre93]

En effet, le point est considéré comme la marque de ponctuation principale pour marquer la limite droite (la fin) d'une phrase. Mais il est clair que le point a d'autres fonctions dans la phrase et peut donc apparaître dans d'autres contextes : à la suite d'un organisateur (1.), après une abréviation (Mme.), etc. De plus, on constate que :

"d'autres signes de ponctuation peuvent marquer la fin d'une phrase : les points de suspension, le point d'interrogation, le point d'exclamation, le point virgule, le double point, mais ces divers signes peuvent aussi se trouver à l'intérieur d'une phrase." [Gre93]

Dans une perspective d'analyse linguistique automatique, l'identification des phrases doit donc tenir compte de différents problèmes liés à l'ambiguïté des marques de ponctuation.

Par ailleurs, le traitement automatique est confronté non pas à des unités linguistiques dans l'absolu mais à des unités inscrites dans des textes. Ainsi, si on tient compte de l'interdépendance entre la structure du document et la structure linguistique au sein d'un texte [Vir89], son interprétation et son analyse ne peuvent pas négliger des phénomènes graphiques, visuels, structuraux etc. qui y sont présents[1].

La prise en compte des caractéristiques physiques d'un texte (sa présentation physique) s'avère nécessaire pour une analyse linguistique précise.

Toutefois, comme nous l'avons montré plus haut, à l'heure actuelle les analyseurs robustes existants modélisent des phénomènes linguistiques apparaissant dans des textes sans pour autant prendre en compte des phénomènes purement textuels. L'unité de base pour de tels systèmes reste la phrase « classique ».

Une telle définition est devenue, pour nous, inappropriée car elle néglige des phénomènes ou des structures qui mettent en jeu des facteurs liés à la présentation visuelle des données linguistiques.

Ainsi, la prise en compte de ces phénomènes nous a amenée à élargir la notion de *phrase*. Selon notre définition, *cf.* 3.2, une phrase est une entité textuelle. La phrase « classique » correspond à cette définition, aussi bien que d'autres unités comme les listes et les titres (que nous avons décrit dans le chapitre 3 de ce rapport). Notre modèle d'analyseur (à la différence de XIP-F) prend alors comme unités de base ces trois types de segment.

Pour identifier correctement les unités intégrant des informations structurelles ou textuelles, nous nous sommes proposé d'enrichir l'analyseur d'un module initial, le but

[1] D'après [Vir89], les contributions de la linguistique à l'étude de structures textuelles génèrent différents types de difficultés qu'il serait dangereux d'ignorer ou de minimiser *("(...) contributions from linguistics to the study of text structures gives rise to several orders of difficulty which it would be dangerous to ignore or to minimize. ")*.

étant de préparer le texte pour obtenir par la suite des traitements syntaxiques plus précis.

Du point de vue de l'architecture de notre modèle d'analyseur, ce module de pré-traitement initial se trouve en amont de tout module linguistique (*cf.* figure 4.1). Son but est d'identifier et de marquer formellement les phrases intégrant des informations structurelles.

Le marquage consiste en l'ajout de balises en tenant compte d'aspects liés à la définition linguistique et structurelle des différents types d'unités : par exemple, à la différence de marquages comme HTML ou XML, le marquage des listes intègre une amorce plus un ensemble d'items, et non des items uniquement.

Par exemple :

(1) *Les derniers pays qui se sont ralliés à l'initiative sont les suivants :*
- *Argentine*
- *Bangladesh*
- *Libéria*
- *Philippines*

Pour cette phrase en exemple, la figure 5.1 montre le marquage spécifique au format HTML. Dans ceci, la balise `<ul>` dénote un début de « liste » et chaque balise `<li>` un item.

```
<p> Les derniers pays qui se sont ralli&eacute;s
&agrave; l'initiative sont les suivants:
<ul>
<li><b>Argentine</b><br>
<li><b>Bangladesh</b><br>
<li><b>Lib&eacute;ria</b><br>
<li><b>Philippines</b></ul></p>
```

FIG. 5.1 – Marquage de type HTML.

Par rapport à ce format, la figure 5.2 montre le balisage produit à la suite de notre pré-traitement.

Nous présentons par la suite la méthodologie que nous avons mise en œuvre pour le marquage de listes (5.2.2) et titres (5.2.3). On notera que les phrases (« classiques ») n'ont pas besoin de balisage spécifique car elles sont déjà bien repérées par l'analyseur[2] .

5.2.2 Marquage de listes

Le balisage des listes est constitué de trois étapes différentes :

– l'identification des amorces (*cf* annexe C),

[2]Nous avons recupéré les règles de XIP-F.

```
<list> Les derniers pays qui se sont ralliés à l'initiative
sont les suivants <sp>:</sp>

<sp> - </sp>  Argentine
<sp> - </sp>  Bangladesh
<sp> - </sp>  Libéria
<sp> - </sp>  Philippines
</list>
```

FIG. 5.2 – Marquage avec notre pré-traitement.

- l'identification des items, et
- le traitement d'autres lignes du texte.

Les autres lignes du fichier correspondent aux structures qui auraient pu être considérées comme des amorces ou des items étant donné leurs caractéristiques (tabulations ou blancs initiaux, présence de deux points, etc.).

Un faisceau d'indices, que nous avons repérés sur corpus, est testé à chaque étape et permet d'identifier les limites droite et gauche de la liste. La partie la plus complexe du traitement est le repérage de la limite droite (fin de liste). Avec notre technique, la précision de repérage des listes et de leur balisage est de 84,5 %, avec un rappel de 94,9 % (*cf.* 5.2.4).

Traitement de l'amorce

Le premier indice permettant de détecter une amorce est la présence d'une ligne qui finit par deux points suivis d'un retour à la ligne. Dans ce cas, la ligne est potentiellement considérée comme un début de liste mais elle ne sera validée qu'après avoir vérifié d'autres indices (par exemple, on ne considère pas comme amorces les lignes avec un seul mot[3]), par exemple :

(2) *Avertissement :*
La synthèse que nous présentons ici s'appuie sur les diagnostics réalisés à la demande d'un certain nombre d'entreprises.

(3) *Note :*
Pour assurer un bon refroidissement de votre moteur, n'hésitez pas à demander à votre garagiste, si votre moteur véhicule est "ancien" (plus de 100000 km) et que vous ne l'ayez pas toujours entretenu, de vérifier le thermostat de chauffage.

Toute ponctuation incluse dans la ligne supposée être une amorce est alors masquée (les deux points deviennent `<sp>:</sp>`, etc.). En effet, ce masquage est nécessaire pour que plus tard le signe de ponctuation ne soit pas identifié (par le module d'identification d'unités de base de l'analyseur) comme une fin de phrase.

[3]Il peut s'agir d'une amorce très simplifiée ; mais d'après ce que nous avons observé dans les corpus, on a préféré traiter ces cas comme des sortes de titres.

La ligne ne sera pas validée comme amorce tant qu'on n'aura pas vérifié la présence d'au moins deux items par la suite. De plus, si une autre structure correspondant à la définition d'amorce est repérée avant d'avoir identifié des items, il est probable qu'il s'agisse d'une liste imbriquée. Dans ce cas complexe, seule la liste intérieure est marquée.

Exemple de liste imbriquée :

(4) *Les États membres prévoient que le responsable du traitement ou son représentant doit fournir à la personne auprès de laquelle il collecte des données la concernant au moins les informations énumérées ci-dessous, sauf si la personne en est déjà informée :*
a) l'identité du responsable du traitement et, le cas échéant, de son représentant ;
b) les finalités du traitement auquel les données sont destinées ;
c) toute information supplémentaire telle que :
- les destinataires ou les catégories de destinataires des données,
- le fait de savoir si la réponse aux questions est obligatoire ou facultative ainsi que les conséquences éventuelles d'un défaut de réponse,
- l'existence d'un droit d'accès aux données la concernant et de rectification de ces données.

Ces cas sont des cas complexes pour le marquage et ne sont pas résolus à l'heure actuelle. Ils ont été recensés et étudiés linguistiquement très en détail par [LGDM+99].

Traitement des items

Pour identifier les items, nous repérons tout d'abord les lignes qui commencent par des espaces, des tabulations, des organisateurs alphanumériques, etc. et on teste aussi leur fin. Étant donné leur nombre, il n'est pas possible de dresser la liste exhaustive des différents cas de figure ; voici les cas principaux :

a. lignes finissant par un point-virgule ou une virgule ;

b. lignes sans signe de ponctuation finale ;

c. lignes finissant par une ponctuation forte.

Le premier cas ne présente pas de problème particulier : tout signe de ponctuation présent dans la ligne est masqué.

Pour le deuxième, s'il n'y a pas de ponctuation finale, on rajoute systématiquement la balise `<sp>.</sp>` pour que l'analyseur puisse identifier plus tard chaque fin d'item.

Finalement, le troisième cas est plus complexe car on est obligé de tester si on se trouve réellement en fin de liste. Voici quelques exemples à la suite de ce traitement :

(5) *Il s'agit d'une simple ceinture ventrale, qui ne peut être efficace que si les conditions suivantes sont remplies* `<sp> :</sp>`

- Le siège est correctement réglé `<sp>,</sp>`
- La ceinture ne frotte pas contre des arêtes vives `<sp>,</sp>`
- La boucle est correctement verrouillée `<sp>,</sp>`
- La sangle est correctement tendue `<sp>.</sp>`

(6) *Après que ce rapport et bien d'autres eurent fait clairement ressortir l'incompétence du ministère, le ministre en personne est allé rendre visite à son personnel à Vancouver* <sp>.</sp> *Voici ce qu'il leur a dit* <sp> :</sp>

L'ultime recours pour conserver et protéger la pêche, c'est la Loi sur les pêches <sp>.</sp> *Si la loi n'est pas respectée il n'y aura pas plus de poisson dans les frayères que ce qui s'y trouve actuellement* <sp>.</sp>

Il devient évident dans la région qu'on ne reçoit pas, aux plus hauts échelons du ministère, cette information sur les questions essentielles de conservation et de protection <sp>.</sp>

La plupart des grandes enquêtes secrètes sur la pêche illégale et les ventes illégales connaissent une fin abrupte <sp>.</sp>

On ne fait pas d'enquêtes complètes sur l'habitat ni d'enquêtes approfondies sur la pêche illégale <sp>.</sp>

Le balisage de la limite droite s'avère le point le plus délicat lors du marquage des listes. Deux possibilités sont à considérer :

A. les lignes précédentes (items) finissent toutes par un signe de ponctuation autre que le point ou bien elles n'ont pas de ponctuation finale, et la ligne actuelle finit par une ponctuation forte ;

B. les lignes précédentes (items) finissent par une ponctuation forte et la ligne potentiellement finale aussi.

Dans le premier cas, on est certain qu'on se trouve à la fin d'une liste. Le balisage est alors complet :

(7) <list> *Cinq classifications internationales différentes ont été successivement établies pour définir les stades cliniques de mélanome* <sp> :</sp>

- *classification de l'Union internationale contre le cancer* <sp>,</sp>
- *classification du MD Anderson* <sp>,</sp>
- *classification à trois stades* <sp>,</sp>
- *classification de TNM de l'American Joint Committee on Cancer* <sp>,</sp>
- *classification conjointe de l'AJCC et de l'UICC* <sp>.</sp>

</list>

Le deuxième cas est source d'ambiguïté car soit la ligne est un autre item (pas la fin de la liste), soit elle est effectivement l'item final de la liste, soit on est déjà sur une ligne hors de la liste qui par ses caractéristiques a été considérée comme un item en raison d'une tabulation ou d'un espace initial.

Les exemples suivants montrent des erreurs de marquage en fin de liste :

(8) <list> *Parallèlement, les entreprises se trouvent confrontées à une forte concurrence, généralement de deux natures différentes* <sp> :</sp>

celle des pays du tiers monde aux très faibles coûts de main d'oeuvre, proposant des prix de 25 à 50% moins cher <sp>.</sp>

celle des autres entreprises aux techniques de production identiques, utilisant les mêmes matières premières, un personnel de niveau de compétence équivalent <sp>.</sp>

Pour faire face à la concurrence et maintenir leur compétitivité sur les prix, certaines entreprises emploient comme moyen privilégié la "délocalisation" d'une partie de leur production <sp>.</sp>

</list>

Si, dans le cas précédent, on peut arriver à éviter le problème en utilisant des indices supplémentaires (minuscule/majuscule des débuts des items, présence de plus d'une ligne blanche), dans d'autres cas l'ambiguïté est plus difficile à lever. Par exemple, dans la phrase (9), les items de la liste commencent par une majuscule et finissent tous par des points, tout comme la phrase suivante. De plus, il n'y a pas d'espace vertical supplémentaire entre le dernier item et la phrase qui suit :

(9) <list> *1) Chez des patients n'ayant pas de tares médicales importantes avec une fonction respiratoire correcte et sans métastases à distance, l'indication peut être difficile dans certains cas* <sp> :</sp>

Le syndrome de Pancoast-Tobias, sans envahissement du corps vertébral <sp>.</sp>

Les tumeurs envahissant la paroi thoracique nécessitent parfois l'utilisation de matériel prothétique <sp>.</sp>

Les tumeurs envahissant la veine cave supérieure sans traduction clinique ou la carène ou l'oreillette gauche <sp>.</sp>

Actuellement aucune étude ne peut démentir cette attitude <sp>.</sp>

</list>

Finalement, étant donné les caractéristiques formelles de quelques parties du texte, certaines structures sont marquées incorrectement comme des listes alors qu'elles ne le sont pas (ceci représente entre 12 % et 15 % d'erreurs sur le total du corpus utilisé pour l'évaluation de ce module, *cf.* 5.2.4) :

(10) <list> *À l'attention du directeur de l'information* <sp> :</sp>

TORONTO, le 27 nov /CNW/ - Selon un nouveau rapport publié aujourd'hui par le procureur général, M Jim Flaherty, et le solliciteur général, M David Tsubouchi, les lois prévoyant la confiscation des biens des criminels se révèlent un élément important dans la lutte contre le crime organisé <sp>.</sp>

Le rapport intitulé Pour que le crime ne paie pas : leçons apprises est fondé sur les travaux d'un sommet international parrainé par le gouvernement de l'Ontario en août dernier, à Toronto <sp>.</sp> *Le sommet a réuni des conférenciers du Canada, des Etats-Unis, de l'Angleterre, du pays de Galles, de l'Irlande et de l'Afrique du Sud* <sp>.</sp>

"Nous avons appris que le crime organisé considère le Canada comme un refuge", a déclaré M Flaherty <sp>.</sp>

</list>

Traitement d'autres lignes

Le module de balisage de listes nettoie enfin les lignes qui présentent des marques `<sp>,</sp>` alors qu'elles n'appartiennent pas à des listes. Ce sont des lignes supposées être des items (ou des amorces) à un moment précis du traitement mais qui se sont révélées fausses une fois le faisceau d'indices vérifié (principalement des lignes avec des tabulations initiales).

5.2.3 Marquage de titres

Le balisage de titres s'avère moins complexe que celui des listes (*cf* annexe C). Du point de vue formel, nous considérons comme titre toute ligne qui a comme fin tout sauf un signe de ponctuation forte ([:] [;] [.] [!] [?]) suivi d'au moins deux retours à la ligne.

Les lignes qui présentent ces caractéristiques subissent le même traitement que les lignes d'une liste en ce qui concerne la ponctuation (toute ponctuation forte est masquée avec les balises `<sp> </sp>` pour éviter des erreurs de l'analyseur lors du découpage en unités de base).

(11) `<tit>` *A* `<sp>.</sp>` *Données statistiques* `</tit>`

(12) `<tit>` *Le curage ganglionnaire* `<sp>;</sp>` *résultats des études contrôlées* `</tit>`

Avec cette technique, la précision de repérage des titres et leur balisage est de 96,8 %, avec un rappel de 96,3 %.

Les quelques erreurs produites correspondent à des erreurs typographiques présentes dans le texte original (généralement des phrases sans ponctuation forte finale) :

(13) `<tit>` *A titre d'exemple, on consomme 5,2l/100 km avec une Mégane 1,9D (pour des vitesses moyennes de 80 km/h sur route avec pointes à 140 km/h) effectuant 40000 km/an soit une économie annuelle de 1600 F environ (il est d'ailleurs possible en se limitant à 90 - 100 km/h maxi de ne consommer que 4,8 l/100 km avec ce véhicule)* `</tit>`

Pour les titres présents dans le texte mais non marqués par le module, il s'agit de lignes finissant par une ponctuation forte comme :

(14) *Et les pneus ?*

(15) *Bush revendique le droit de polluer la planète.*

Ce type de structure n'est pas considéré comme titre dans notre approche. Il s'agit d'une unité de base « classique ».

5.2.4 Évaluation

L'évaluation de notre module de pré-traitement a été faite sur un ensemble de corpus variés extraits du Web (*cf.* annexe A). La surface représentée par les listes et les titres (c'est-à-dire le pourcentage de mots du corpus appartenant à ces structures) est de 13,55 %. Le balisage présente une moyenne des taux de précision et de rappel (mesure

F1) de 89,4 % pour les listes (précision à 84,5 %, rappel à 94,9 %) et de 96,6 % pour les titres (précision à 96,8 %, rappel à 96,3 %).

À partir de cette évaluation, nous avons relevé deux sortes de problèmes. D'une part, les conventions typographiques établies ne sont pas toujours respectées dans les textes originaux (manque de ponctuation) ce qui complique le balisage. D'autre part, une analyse plus fine permettrait peut être de couvrir des cas qui ne sont pas bien traités jusqu'à présent (i.e. listes imbriquées).

Malgré ces problèmes, nos deux modules constituent un filtre indispensable avant l'application de l'analyseur car ils permettent le repérage de phrases ayant trait à des aspects liés à la ponctuation et à la structure du document.

5.3 La Grammaire Noyau

Ainsi que nous l'avons expliqué dans le chapitre précédent, et à la différence de XIP-F, le modèle d'analyseur que nous proposons est fondé sur une division de l'analyse linguistique en deux étapes, définies à partir des caractéristiques linguistiques et structurelles de chaque unité en entrée (phrases N1 et phrases N2, voir figure 4.1).

Une grammaire noyau formalise les caractéristiques des phrases que nous appelons de premier niveau (N1). Comme nous le montrons plus loin, en général, les phrases de ce type ont une moyenne de 18 mots et représentent environ 20 % des phrases d'un ensemble de corpus variés (*cf.* annexe A).

5.3.1 Fondements théoriques

La notion de « noyau » en linguistique est apparue dans les toutes premières études de grammaire classique. Apollonius Dyscolus, l'auteur du Péri Syntaxéôs (IIe siècle ap. J.-C.) [Lal97], montre que l'énoncé se construit autour d'un noyau nom-verbe (ou noyau nom-verbe-nom) : tous les autres éléments peuvent être enlevés sans que l'énoncé soit incomplet.

Les phrases constituées d'un ou deux éléments linguistiques autour d'un verbe principal (conjugué) sont alors des « phrases noyau ». Cette définition se rapproche de la notion de « phrase minimale » [Dub69] :

"Unité élémentaire de l'énonciation (...) formée de deux constituants, l'un est un syntagme nominal (SN), appelé sujet, et l'autre un syntagme verbal (SV), qui se voit accorder le statut de prédicat."

Pour [Dub69], on distingue :

"deux schémas disctincts selon que SV est constitué d'un verbe de type avoir et d'un SN complément ou que SV est formé d'un verbe du type être et d'un adjectif-participe. Selon l'option, on obtient deux phrases nucléaires ou phrases-noyaux."

Par exemple :

(16) *L'enfant lit un livre.*

(17) *L'enfant est malade.*

En linguistique « classique » ces deux phrases sont formées de deux constituants ou syntagmes (définis comme *"un groupe de mots formant une unité à l'intérieur de la phrase, un groupe ayant une fonction dans la phrase."*, [Gre93]).

Dans notre approche, nous avons formalisé la notion de « phrase noyau », étendue par rapport à la notion de phrase minimale, à partir de l'étude de corpus de textes hétérogènes. L'ensemble de règles qui décrivent ce type de phrases constitue notre « grammaire noyau ».

Il existe, dans la littérature, d'autres grammaires dites noyau, par exemple GNF (Grammaire Noyau du Français) [PS96], leur but étant de décrire les constructions fondamentales d'une langue donnée.

Dans notre approche, la grammaire noyau (GN) a comme objectif le découpage en syntagmes de base de toutes les phrases repérées dans un texte en entrée et, par la suite, l'identification des constructions fondamentales, c'est-à-dire, celles qui ne requièrent pas des traitements syntaxiques spécialisés.

5.3.2 Description générale de la grammaire noyau

Concrètement, la grammaire noyau est constituée d'un ensemble de règles[4] permettant :

a) le découpage en segments de base (syntagmes noyau ou *chunks*),

b) le découpage en segments additionnels (groupes de syntagmes noyau et propositions subordonnées),

c) le marquage de phrases de premier niveau.

Le texte en entrée de la grammaire présente les marques des transformations du prétraitement : il contient des balises de marquage d'unités de base pour les listes et les titres. De même, chaque unité lexicale présente une étiquette morphologique désambiguïsée ainsi qu'un ensemble de traits.

Les règles de la grammaire noyau utilisent ces informations pour effectuer une première analyse syntaxique qui permet d'identifier les syntagmes noyau de base (comme dans la plupart des analyseurs robustes existants). Par la suite, des règles plus complexes s'appliquent pour trier les unités considérées noyau (N1) des unités requérant une analyse structurelle plus approfondie (phrases N2).

a) Segments de base

Le découpage en segments de base permet d'obtenir un marquage en syntagmes noyau. Nous considérons comme segments de base les six syntagmes noyau suivants :

- NP (syntagme nominal)
- AP (syntagme adjectif)
- PP (syntagme prépositionnel)

[4] Un échantillon de ces règles se trouve dans l'annexe D.1.

- IV (syntagme verbal infinitif)
- FV (syntagme verbal conjugué)
- GV (syntagme verbal participe présent)

À la différence de la grammaire de *chunking* de XIP-F pré-existant (*cf.* annexe B.2), nous ne définissons pas de proposition noyau (`SC`) et les débuts de groupe et propositions subordonnées (`BG` et `SBC`, respectivement) sont analysés dans une étape postérieure. Le nœud maximal de l'arbre a, dans notre approche, l'étiquette `MAX`.

Il existe à l'heure actuelle 40 règles pour la définition des syntagmes noyau de base[5] (2 pour `AP`, 23 pour `NP`, 7 pour `PP`, 2 pour `FV`, 4 pour `IV` et 2 pour `GV`), leur application se faisant toujours de façon incrémentale.

Toute phrase du texte en entrée est marquée selon les critères définis par ces règles. Voici deux exemples d'analyse par la grammaire noyau correspondant à ce que la grammaire identifiera plus tard comme une phrase N1 et une phrase N2, respectivement[6] :

(18) *Les dispositions législatives ne peuvent pas l' emporter sur la Constitution .*

```
5788>MAX{NP{Les dispositions} AP{législatives} FV{ne peuvent} pas
IV{l' emporter} PP{sur NP{la Constitution}} .}
```

(19) *Par ailleurs , les nouveaux redevables sont soumis à des règles particulières concernant :*
- l' utilisation d'un prorata provisoire pendant leur première année d'activité ;
- la déduction d'un "crédit de départ" de TVA .

```
395>MAX{<list> Par ailleurs , NP{les nouveaux} AP{redevables} FV{sont
soumis} PP{à NP{des règles}} AP{particulières}} concernant <sp>:</sp>
- NP{l' utilisation} PP{d' NP{un prorata}} AP{provisoire} PP{pendant
NP{leur AP{première} année}} PP{d' NP{activité}} <sp>;</sp>
- NP{la déduction} d' un " NP{crédit} PP{de NP{départ}} " NP{de TVA}
<sp>.</sp> </list>}
```

Parfois, des erreurs d'analyse peuvent se produire. Dans la deuxième phrase en exemple, c'est le cas de *d'un "crédit de départ"*. L'erreur d'analyse est provoquée par la présence des guillemets qui ne sont pas modélisés dans les règles des syntagmes de base.

Ces types d'erreur seront résolus plus tard par la grammaire spécialisée dans le traitement de marques de ponctuation particulières (guillemets et parenthèses, voir chapitre 6, section 6.3).

b) Segments additionnels

Le découpage en segments additionnels que nous avons modélisé consiste, tout d'abord, à identifier et à marquer des propositions subordonnées ; par la suite les différents syntagmes obtenus jusqu'ici sont regroupés en trois sortes de structures autour du `NP` et `FV`

[5] Elles sont basées sur le travail de Salah Aït-Mokhtar pour la grammaire de XIP-F.
[6] Le chiffre initial correspond au numéro de la phrase dans le corpus.

principaux. Le regroupement de syntagmes permet d'identifier ce que nous appelons des « expansions » (syntagme ou groupe de syntagmes dépendant d'une tête syntaxique, *cf.* 3.2).

Nous avons défini trois types d'expansions :

`ANP` (groupe antérieur au `NP` sujet) ;

`PNP` (groupe postérieur au `NP` sujet) ;

`PFV` (groupe postérieur au `FV`) .

Ces groupes de syntagmes permettent de contrôler la nature et l'organisation des éléments dans une phrase autour des deux éléments traditionnellement considérés comme composant le noyau. Leur analyse permettra de distinguer les phrases N1 des phrases N2 lors de l'étape suivante de l'analyse.

ANP

Il s'agit d'une expansion précédant le `NP` principal d'une phrase (ce `NP` sera plus tard marqué comme sujet). Nous avons écrit 26 règles permettant de définir la nature des éléments intégrant cette structure. Par exemple :

```
9>ANP[start:+] -> pp[start:+].

9>ANP[start:+] -> ap[start:+], adv*, pp*, punct[last:+,form:fcm].
```

La première règle de l'exemple précédent définit un `ANP` comme un syntagme prépositionnel ayant le trait `start`[7] (ce trait est alors hérité par l'expansion). Dans la deuxième règle, le `ANP` défini a comme premier élément un syntagme adjectival, il contient zéro ou plusieurs adverbes et/ou syntagmes prépositionnels et finit par une ponctuation de type virgule.

Exemple :

`ANP{PP{`En `NP{`1995 `}}` ,`}` *les exportations de produits agroalimentaires ont atteint un total de 92 millions de dollars.*

Les `ANP` plus complexes intègrent au maximum deux virgules et une proposition subordonnée.

PNP

Les PNP sont des groupes des syntagmes postérieurs au `NP` (sujet) et antérieurs au `FV` principal.

Exemple pour la phrase précédente :

En 1995, les exportations `PNP{PP{`de `NP{`produits`}}` `AP{`agroalimentaires`}}` *ont atteint un total de 92 millions de dollars.*

Nous avons écrit un total de 25 règles permettant de décrire la nature des éléments

[7] Pour rappel, le trait `start` dénote le premier élément de l'unité de base (la phrase en entrée) ; le trait `last` marque le dernier élément de l'unité définie, dans ce cas le `ANP`.

intégrant les PNP. Par exemple :

```
8>PNP -> |np| pp[first:+], pp*, adv*, punct[form:fcm], (punct[form:fcm])
|fv[verb:+,fin:+]|.
```

Les règles permettent de définir clairement les contextes gauche (NP) et droit (FV). La règle donnée en exemple définit un PNP consitué d'un syntagme prépositionnel initial, de zéro ou plusieurs syntagmes prépositionnels et/ou adverbiaux et d'une (ou deux) virgules[8]. Exemple :

À l' été de 1994, la gestion PNP{PP{de NP{la pêche}} PP{au NP{saumon}} PP{du NP{Pacifique}} , PP{par NP{le ministère}} ,} *était le chaos le plus total.*

PFV

Finalement, une expansion de type PFV regroupe des syntagmes postérieurs au verbe principal. Nous avons écrit 16 règles décrivant l'ensemble des possibilités pour cette structure. Par exemple :

```
11>PFV -> |fv| np, adv*, negpas*, np*, ap*, pp*, gv*, iv*, verb[part:+,pas:+]*,
(punct[form:fcm]), (punct[form:fcm]), (sbc), (bg) |~fv|.
```

Cette règle définit un PFV constitué d'un syntagme nominal obligatoire, de zéro ou plusieurs autres types de syntagmes (NPs, PPs, etc.) ou unités lexicales (adverbes, négations) et éventuellement d'un maximum de deux virgules et une proposition subordonnée. Comme limite droite il ne peut pas y avoir un syntagme verbal conjugué.

Les deux analyses suivantes montrent la sortie des traitements définis jusqu'ici sur les deux phrases N1 et N2 données en exemple plus haut (18) et (19) :

```
5788>MAX{NP{Les dispositions} PNP{AP{législatives}} FV{ne peuvent}
PFV{pas IV{l'emporter} PP{sur NP{la Constitution}}} .}

395>MAX{<list> Par ailleurs , NP{les nouveaux} PNP{AP{redevables}}
FV{sont soumis} PP{à NP{des règles}} AP{particulières} concernant
<sp>:</sp> - NP{l' utilisation} PP{d' NP{un prorata}} AP{provisoire}
PP{pendant NP{leur AP{première} année}} PP{d' NP{activité}} <sp>;</sp>
- NP{la déduction} d' un " NP{crédit} PP{de NP{départ}} " NP{de TVA}
<sp>.</sp> </list>}
```

On observe que pour la deuxième phrase les structures ANP et PFV n'ont pas été identifiées. Des éléments différents de ceux acceptés par ces structures sont présents et empêchent leur reconnaissance. Par exemple, les balises provenant du pré-traitement ou bien certaines marques de ponctuation comme les guillemets, bloquent l'analyse de ces structures.

[8]Pour rappel, le trait first dénote le premier élément de l'unité définie, ici le PP.

c) Marquage de phrases N1

L'identification de syntagmes et expansions réalisée dans les étapes précédentes, permet dans ce dernier stade d'analyse de la grammaire noyau d'identifier les phrases noyau.

Le segment (nœud non lexical dans l'arbre syntaxique) correspondant à une phrase de type N1 est représenté par `S`. Nous avons écrit un total de 7 règles permettant de définir ce type de phrases ; chaque règle inclut les possibilités définies par la règle précédente (en raison de l'application successive des règles) :

R1 Un `NP` et un `FV` sans expansions (*phrase minimale*) ou bien un seul segment de base (et un ou plusieurs adverbes) intégrant le `PFV`.

(20) *Les gains ont été limités.*

```
56>MAX{S{NP{Les gains} FV{ont été limités} .}}
```

(21) *Plusieurs pays ont adopté cette politique.*

```
160>MAX{S{NP{Plusieurs pays} FV{ont adopté} PFV{NP{cette politique}} .}}
```

R2 Expansions antérieures et/ou postérieures au NP sujet.

(22) *Toutefois , une grande variété de produits alimentaires doivent être importés.*

```
29>MAX{S{ANP{Toutefois ,} NP{une AP{grande} variété} PNP{PP{de
NP{produits}} AP{alimentaires}} FV{doivent} PFV{IV{être importés}} .}}
```

R3 Deux segments de base (et un ou plusieurs adverbes) intégrant le `PFV`.

(23) *La Bourse de Paris a terminé sur une note négative.*

```
2347>MAX{S{NP{La Bourse} PNP{PP{de NP{Paris}}} FV{a terminé}
PFV{ PP{sur NP{une note}} AP{négative}} .}}
```

R4 n segments de base (et un ou plusieurs adverbes) intégrant le `PFV`.

(24) *Un cliquet maintiendra la fenêtre ouverte dans la position voulue.*

```
2987>MAX{S{NP{Un cliquet} FV{maintiendra} PFV{NP{la fenêtre}
AP{ouverte} PP{dans NP{la position}} AP{voulue}} .}}
```

R5 Un segment additionnel (proposition subordonnée) intégrant l'`ANP`, le `PNP` ou le `PFV` avec éventuellement une virgule.

(25) *Si ces raies sont trop nombreuses ou trop larges, elles constituent un à trois pics calculables au moyen de formules compactes.*

```
3401>MAX{S{ANP{SBC{BG{Si} NP{ces raies} FV{sont}} AP{trop nombreuses
ou trop larges} ,} NP{elles} FV{constituent} PFV{NP{NUM{un à trois} pics}
AP{calculables} PP{au moyen de NP{formules}} AP{compactes}} .}}}
```

(26) *La thoracoscopie permet de réaliser une résection en coin du parenchyme pulmonaire où siègent le ou les nodules pulmonaires.*

```
4353>MAX{S{NP{La thoracoscopie} FV{permet} PFV{IV{de réaliser}
NP{une résection} PP{en NP{coin}} PP{du NP{parenchyme}} AP{pulmonaire}
SBC{BG{où} FV{siègent}} NP{le ou les nodules} AP{pulmonaires}} .}}
```

R6 Deux segments additionnels (propositions subordonnées) intégrant l'`ANP`, le `PNP` ou le `PFV` avec éventuellement deux virgules.

(27) *On estime que sa population est à près de 45,1 millions de personnes et que le taux de natalité y est de plus de 2,5 %.*

```
61>MAX{S{NP{On} FV{estime} PFV{SBC{BG{que} NP{sa population} FV{est}}
PP{à près de NP{45,1 millions}} PP{de NP{personnes}} SBC{BG{et que}
NP{le taux de natalité} FV{y est}} de plus PP{de NP{NOUN{2,5 \%}}}} .}}
```

(28) *Les fonds seront attribués aux banques sud-africaines qui les prêteront, à leur tour, à des acheteurs sud-africains.*

```
139>MAX{S{NP{Les fonds} FV{seront attribués} PFV{PP{aux NP{banques}}
AP{sud-africaines} SBC{BG{qui} FV{les prêteront}} , PP{à NP{leur tour}} ,
PP{à NP{des acheteurs}} AP{sud-africains}} .}}
```

R7 Coordinations non ambiguës (par exemple, dans des configurations comme `[NP (AP) coord NP (AP) FV]` ou bien `[FV (PFV) coord FV (PFV)]` où `PFV` ont la même structure syntagmatique).

(29) *Le gouvernement sud-africain a cependant commencé à démanteler les barrières commerciales et s'est engagé à instaurer un système commercial ouvert.*

```
172>MAX{S{NP{Le gouvernement} PNP{AP{sud-africain}} FV{a cependant
commencé} PFV{IV{à démanteler} NP{les barrières} AP{commerciales}}
et FV{s' est engagé} PFV{IV{à instaurer} NP{un système} AP{commercial}
AP{ouvert} .}}
```

Le tableau suivant résume les caractéristiques principales de chaque règle `S` de la grammaire et donne une idée de la taille moyenne des phrases (en mots) correspondant à chaque définition :

Règle	*Caractéristiques*	*Longueur moyenne des phrases*
R1	0 ou 1 constituants dans `PFV`	5,7 mots/phrase
R2	`ANP` et `PNP`	9,6 mots/phrase
R3	2 constituants de base dans `PFV`	10,8 mots/phrase
R4	n constituants de base dans `PFV`	14,9 mots/phrase
R5	1 `SBC` et 1 virg. dans `PNP` et `PFV`	17 mots/phrase
R6	2 `SBC` et 2 virg. dans `PFV`	17,9 mots/phrase
R7	coordinations non ambiguës	18,5 mots/phrase

Les phrases du corpus ayant été identifiées comme des phrases de premier niveau correspondent aux caractéristiques décrites par les sept règles précédentes.

À ce stade de l'analyse, ces phrases N1 sont dirigées vers le module d'extraction de dépendances de base. Les autres (phrases de deuxième niveau) sont dirigées vers les modules de traitement syntaxique spécialisé (analyse structurelle et extraction de dépendances spécifiques).

5.3.3 Extraction de dépendances de base

Comme nous l'avons mentionné au chapitre précédent, l'extraction de dépendances syntaxiques constitue l'objectif final de l'analyseur. Les règles définissant les dépendances de base sont regroupées dans un module indépendant à la grammaire noyau (voir figure 4.1). Cependant, nous décrivons ici ce module de façon succinte car il complète l'analyse syntaxique pour les phrases de niveau 1.

Dans notre approche, nous avons défini 12 relations syntaxiques de base[9]. Elles sont extraites des phrases N1 marquées par la grammaire noyau et également des phrases N2 une fois que le marquage de segments et l'extraction de dépendances spécialisées a été mise en œuvre.

Pour rappel, une dépendance est une structure de deux éléments (parfois trois) de type `nom(x,y)` où `nom` est le nom de la dépendance, `x` est l'élément noyau ou « tête » syntaxique et `y` l'élément dépendant.

La liste suivante décrit un échantillon de dépendances de base. Nous renvoyons le lecteur à l'annexe D.4 pour une liste détaillée avec des exemples :

`SVO(x,y,z)` structure argumentale de base de la phrase

`SUBJ(x,y)` sujet d'un `FV`, `IV` ou `GV`

`VPP(x,y,z)` complément prépositionnel d'un `FV`

`NADJ(x,y)` complément adjectival d'un `NP`

Dans le cadre de notre travail sur le traitement des ambiguïtés structurelles, nous avons modélisé notre propre grammaire pour l'extraction de dépendances liées au traitement du rattachement prépositionnel (*cf.* chapitres 7 et 8).

Le résultat final de l'analyse pour la phrase N1 donnée en exemple plus haut est le suivant :

[9]Leur définition s'appuie sur le travail réalisé par Jean-Pierre Chanod dans le cadre de la grammaire pour XIP-F.

```
SUBJ(peuvent,disposition)
VPP(emporter,sur,Constitution)
NADJ(disposition,législatives)

5788>MAX{S{NP{Les dispositions} PNP{AP{législatives}} FV{ne peuvent}
PFV{pas IV{l'emporter} PP{sur NP{la Constitution}}} .}}
```

Dans cette phrase, on peut remarquer deux expansions, l'une postérieure au syntagme nominal sujet, l'autre postérieure au verbe principal. L'ensemble des structures analysées correspond à une phrase de premier niveau marquée comme `S`.

5.3.4 Évaluation

Le corpus utilisé pour évaluer l'ensemble de règles appartenant à la grammaire noyau est constitué d'un ensemble de textes de domaines et de genres différents d'environ 87.000 mots (4.000 phrases) :

- Journaux généraux *Le Monde* (16.876 mots) et *Libération* (4.718) ;
- Journal de finances *Les Echos* (13.755) ;
- Rapports juridiques (11.622) ;
- Rapports scientifiques de médécine (14.395) et de physique et chimie (2.575) ;
- Manuel technique (9.385) et guide d'entretien d'une voiture (14.716).

L'ensemble de ces corpus ont été extraits du Web en juilllet 2000. L'évaluation des différentes versions de la grammaire noyau a été faite manuellement sur 102 phrases sélectionnées au hasard.

Marquage structurel

La table suivante montre les résultats concernant la couverture et la précision du marquage des phrases de premier niveau (chiffres obtenus après une évaluation manuelle des corpus mentionnés plus haut). Nous définissons ici la « couverture » comme le pourcentage de phrases du corpus correspondant à la définition de phrase noyau (N1). La précision correspond au pourcentage de phrases du corpus –ayant une structure noyau– qui ont été bien marquées en tant que phrase noyau (nœud `S`) par l'analyseur.

Type de règle	*Couverture*	*Précision*
R1	0 % à 0,5 %	100 %
R2	0,5 % à 1,5 %	100 %
R3	1,5 % à 3,5 %	100 %
R4	3,5 % à 10 %	100 %
R5	10 % à 15 %	98,1 %
R6	15 % à 17,5 %	96,4 %
R7	17,5 % à 20 %	**96,1 %**

La couverture estimée pour la dernière règle (qui inclut les heuristiques définies par les autres règles) correspond à la couverture finale de la grammaire noyau. Cela implique

que, quel que soit le corpus à analyser par notre système, environ 20 % des phrases corresponderont à la notion de phrase de premier niveau et seront traitées avec 96 % de précision lors du marquage structurel.

Extraction de dépendances

Dans la table suivante, la précision est le nombre de relations correctes sur le nombre de relations extraites. Sept dépendances ont été évaluées :

Type de phrase/règle	*Précision*
SVO	100 %
VN / VADJ	100 %
SUBJ	99,1 %
OBJ	95,6 %
VPP	78,5 %
NPP	77,9 %

La précision concernant l'extraction de dépendances est fortement liée à la nature de la relation syntaxique. Ainsi pour les relations de base, c'est-à-dire sujet, objet et modifieurs verbaux, la précision se situe autour de 95 %. En revanche, d'autres relations plus complexes comme le rattachement prépositionnel, ont des seuils beaucoup plus bas du fait des problèmes d'ambiguïté structurelle (qui ne sont pas traitables de façon efficace uniquement avec des informations syntaxiques, *cf.* chapitres 7 et 8).

5.4 Résumé

Nous avons vu, dans ce chapitre, deux aspects fondamentaux caractérisant le modèle d'analyseur que nous proposons, à savoir, un module de prétraitement et une grammaire noyau.

D'une part, nous avons décrit en détail le module de prétraitement linguistique. L'objectif de ce module est le repérage d'unités de base (phrases en entrée) ayant des caractéristiques formelles particulières requérant une analyse préalable aux traitements de l'analyseur. L'application d'un tel module permet une meilleure prise en compte de structures liées à des notions au delà de la linguistique classique : notion de document, de visualisation de l'information textuelle, etc. Listes et titres seront mieux analysés ultérieurement par l'analyseur grâce aux traitement fournis par ce module.

D'autre part, nous avons présenté en détail la grammaire noyau de notre système ainsi qu'un module de dépendances de base. La grammaire noyau est un module constitué de règles de grammaire identifiant des structures avec un taux de précision élévé. Ce module permet aussi de distinguer deux types de phrases selon les structures, unités particulières, etc. présentes dans chaque cas.

En effet, les phrases présentant des caractéristiques bien modélisées par les règles de la grammaire noyau sont considérées comme des phrases de premier niveau (N1). Elles constituent environ 20 % de tout type de corpus tout venant. Notre analyseur est ainsi capable de produire une analyse fine (environ 95 % de précision) pour 20 % des phrases d'un corpus quel que soit son domaine.

Le reste des phrases, que nous appelons de deuxième niveau (N2), contient des structures ou des unités spécifiques (ponctuation, tabulations, etc.) qui demandent des traitements syntaxiques particuliers.

Nous allons voir, dans le chapitre suivant, l'ensemble des grammaires créées pour traiter ces différents phénomènes et constituant le deuxième niveau d'analyse de notre système. Ce deuxième niveau s'avère plus complexe car il a pour objectif de traiter des phrases tout venant quelles que soient leurs caractéristiques structurelles.

Chapitre 6

Deuxième niveau d'analyse

6.1 Introduction

Comme nous l'avons décrit précédemment, le modèle d'analyseur que nous proposons procède à une analyse linguistique en deux étapes. La première étape, décrite au chapitre précédent, permet le traitement des phrases de premier niveau (N1) grâce à l'application d'un ensemble de règles constituant une grammaire noyau. La deuxième étape est dédiée à des phrases de deuxième niveau (N2) qui, de par leurs caratéristiques linguistiques et/ou structurelles, demandent un traitement syntaxique particulier.

Dans ce chapitre, nous montrons les différentes grammaires que nous avons créées pour modéliser certains des phénomènes considérés de deuxième niveau, à savoir, des phénomènes liés :

- à la ponctuation (guillemets et parenthèses) ;
- à la structure des documents (titres) ;
- à la ponctuation et à la structure des documents (listes et énumérations) ;

Les phrases sans guillemets ni parenthèses n'étant pas titres, ni listes, ni énumérations et qui ne respectent pas les critères définis pour les phrases N1, sont aussi considérées de deuxième niveau (elles constituent environ un tiers des corpus, *cf.* section 6.7). Cependant, les phénomènes particuliers qu'elles présentent (par exemple, des coordinations ambiguës, des appositions, des propositions interrogatives, etc.) n'ont pas été pris en compte dans le cadre de notre travail. Ces phrases sont analysées par l'ensemble de nos grammaires, mais seulement en ce qui concerne le marquage de syntagmes noyau et l'extraction de dépendances de base.

En tenant compte de ces constats, le chapitre se décompose en plusieurs sections. La première présente une description générale des grammaires spécialisées. Les notions de *chunking* et d'extraction de dépendances sont abordées de façon générale. Les sections suivantes traitent en détail chacune des grammaires implémentées dans notre système.

6.2 Description générale des grammaires

Nos grammaires spécialisées sont des ensembles de règles adaptées au traitement de quelques phénomènes particuliers considérés complexes. Le traitement syntaxique consiste en un découpage en segments et en une extraction de dépendances syntaxiques.

Les phrases que nous considérons comme de deuxième niveau sont ainsi traitées par les modules spécialisés correspondant aux phénomènes qu'ils présentent. Les modules grammaticaux sont activés uniquement en fonction de l'identification des différents phénomènes. L'objectif de ces modules est double : marquer les segments concernant des phénomènes complexes (et, accessoirement, corriger des erreurs éventuelles produites lors du marquage des segments de base) et extraire les liens de dépendance entre les têtes syntaxiques des phénomènes traités.

6.2.1 Marquage de phrases N2

Les segments spécialisés que nous avons définis pour les phénomènes complexes des phrases du deuxième niveau sont les suivants :

- `GM` (structure entre guillemets)
- `PR` (structure entre parenthèses)
- `SP` (séparateur dans les cas de listes)
- `AMR` (amorce d'une liste)
- `IT` (item d'une liste)
- `ENM` (énumération)

Ces segments sont des super-structures pouvant contenir un seul élément (un chiffre dans le cas des séparateurs), des syntagmes noyau, des phrases entières, etc. L'annexe D.2 présente quelques échantillons des règles de grammaire spécialisées.

Voici quelques exemples à la sortie des traitements effectués par ces modules :

(1) *La BNP négociera le rachat des filiales africaines pour ensuite les fusionner avec les banques de son réseau africain, les BICI* **PR[***(Banques internationales pour le commerce et l'industrie)* **]**.

(2) **GM[***"Il ne faut pas s'attendre à un net ralentissement des prix avant 1991"***]**, *assure une grande banque.*

(3) **AMR[** *L'introduction des technologies nouvelles que nous avons observées, dans toutes les entreprises confondues, peut se résumer comme suit :* **]**

IT[SP[-] *Automatisation de l'opération d'enformage* **]**
IT[SP[-] *Coupe à commande numérique* **]**
IT[SP[-] *Alimentation, coloration automatique* **]**
IT[SP[-] *Mise en place d'une GPAO* **]**

(4) *Le présent article s'appuie sur un certain nombre d'interventions, diagnostics ou études réalisés par le réseau Anact, dans les services publics ou non marchands* **PR[ENM[** *(CAF, collectivités territoriales, travail social, hôpital)***]** **]**.

Nous décrivons en détail chacun de ces segments dans les sections correspondant à chaque phénomène.

6.2.2 Correction d'erreurs de chunking

Nos grammaires spécialisées permettent aussi la correction de certaines erreurs de segmentation produites par la non prise en compte de phénomènes complexes au moment de l'analyse des syntagmes noyau de base. C'est le cas, par exemple, de syntagmes nominaux mal segmentés à cause de la présence de guillemets.

Ainsi pour la phrase suivante :

(5) *Les "Business Divisions" ont la responsabilité complète d'un produit de A à Z.*

Une première analyse en syntagmes noyaux donne le découpage suivant :

```
43>MAX{Les " NP{Business Divisions} "  FV{ont}
NP{la responsabilité} AP{complète} PP{d' NP{un
produit}} de A à Z .}
```

Après l'intervention de la grammaire spécialisée dans le traitement de la ponctuation, l'erreur initiale (l'article qui n'a pas été pris en compte à cause des guillemets), est intégré au syntagme :

```
43>MAX{NP{Les GM{" NP{Business Divisions} "}}
FV{ont} NP{la responsabilité} AP{complète} PP{d'
NP{un produit}} de A à Z .}
```

6.2.3 Extraction des dépendances

Après le repérage des syntagmes noyau et d'autres constituants spécialisés, l'analyseur calcule les dépendances syntaxiques entre les têtes de ces syntagmes et constituants. L'annexe D.3 présente quelques échantillons des règles d'extraction de dépendances spécialisées. Nous en donnons les caractéristiques principales dans les sections suivantes correspondant à chaque phénomène traité (guillemets, parenthèses, titres, listes et énumérations).

6.3 Traitement de la ponctuation

Comme nous l'avons constaté au chapitre 3, les traitements spécifiques des marques de ponctuation en linguistique informatique sont peu nombreux. Cela est surprenant si on tient compte du nombre élevé de phénomènes liés à la ponctuation présents dans un texte.

Dans notre approche, nous avons choisi de prendre en compte la ponctuation, directement dans le cas de structures délimitées par guillements et parenthèses (phénomènes de clôture) et indirectement dans le cas de signes de ponctuation intégrant d'autres structures plus complexes (par exemple, les points et virgule [;] dans des listes).

La grammaire créée pour le traitement syntaxique des phénomènes de clôture est l'objet de cette section. Ce module fournit un marquage spécifique des structures liées à la ponctuation et permet l'extraction de dépendances entre leurs têtes et les éléments de la phrase dont ils dépendent.

6.3.1 Segments spécialisés

Suite à notre étude décrite au chapitre 3, nous avons considéré tous les éléments délimités par des marques de clôture comme des unités entières. Pour ce faire, nous avons défini les segments `GM` (guillemet) et `PR` (parenthèse). Sept règles de notre grammaire permettent de prendre en compte les différents cas de figure.

La règle suivante, par exemple, marque comme unité `PR` tout ce qui se trouve à l'intérieur d'une parenthèse (syntagmes noyaux inclus) :

```
10>PR = punct[form:fopar], ?*, punct[form:fcpar].
```

Par exemple :

(6) *Si tout se passe comme prévu, c'est Michel Jalmain (CFDT) qui doit succéder à Denis Gautier-Sauvagnac (Medef).*

```
631>MAX{SBC{BG{Si} NP{tout} FV{se passe}} comme
prévu , NP{c'} FV{est} NP{Michel Jalmain} PR{(
NP{CFDT} )} SBC{BG{qui} FV{doit}} IV{succéder}
 PP{à NP{Denis Gautier}} PR{( NP{Medef} )} .}
```

(7) *Après oxydation de l'aluminium, ce catalyseur possède une surface spécifique de 38 m2/g (de même ordre de grandeur que celle des catalyseurs commerciaux).*

```
3945>MAX{PP{Après NP{oxydation}} PP{de NP{l'
aluminium}} , NP{ce catalyseur} FV{possède}
NP{une surface} AP{spécifique} PP{de NP{NOUN
{38 m2}}} / NP{g} PR{( de même NP{ordre} PP{de
NP{grandeur}} que NP{celle} PP{des NP{catalyseurs}}
AP{commerciaux} )} .}
```

Dans le cas des guillemets, la règle suivante marque comme `GM` tout ce qui se trouve en début de phrase entre guillemets et suivi d'une virgule avec un verbe conjugué comme contexte droit[1].

[1] L'expression entre | | exprime ce contexte

```
24>GM = ?[form:~fquotes,start:+], ?*[form:~fquotes], punct[form:fquotes]
 |punct[form:fcm], fv|.
```

Par exemple :

(8) *« Les pouvoirs publics ont perdu le contrôle de l'économie à cause de l'absence d'un taux d'épargne décent », a-t-il constaté.*

```
GM{<< NP{Les pouvoirs publics} FV{ont perdu} NP{le contrôle} PP{de
NP{l'économie}} PP{à cause de NP{l' absence}} PP{d' NP{un taux}}
PP{d' NP{épargne}} AP{décent} >>} , FV{a NP{-t-il} constaté}.
```

6.3.2 Dépendances

Une dépendance ou fonction syntaxique rattache un mot dans le contexte duquel il figure [Mar85]. Dans ce sens « fonction » s'oppose à « nature » (syntaxe *vs* morphologie) : un mot en contexte dans une phrase ou syntagme est différent d'un mot isolé qui garde sa nature mais perd sa fonction.

A. Martinet[2] définit trois types de rapports fondamentaux entre des éléments (structures syntaxiques) [Mar85] présentés dans la figure 6.1.

(1) $\mathcal{A}$ et $\mathcal{B}$ n'existent pas l'un sans l'autre : $\mathcal{A}$ suppose $\mathcal{B}$ et $\mathcal{B}$ suppose $\mathcal{A}$ $(\mathcal{A} \longleftrightarrow \mathcal{B})$;

(2) $\mathcal{A}$ peut exister sans $\mathcal{B}$, mais $\mathcal{B}$ ne peut pas exister sans $\mathcal{A}$ $(\mathcal{A} \longleftarrow \mathcal{B})$;

(3) $\mathcal{A}$ et $\mathcal{B}$ coexistent sans se conditionner, sorte de co-présence $(\mathcal{A} — \mathcal{B})$.

FIG. 6.1 – Typologie de relations entre des structures syntaxiques.

Nous tenons compte de cette classification pour caractériser les fonctions syntaxiques qui concernent les phénomènes de clôture[3].

La relation (1) (appellée *nexus* par [Jes69]) est la liaison directe entre le verbe et le sujet, le verbe et son objet ou le verbe et un élément prédicatif (éléments régis obligatoires). Nous l'appelons *dépendance directe*. Ce type de dépendance concerne principalement les guillemets. Les parenthèses ne sont possibles que quand le verbe est intransitif ou bien lorsqu'il a déjà un objet. La justification est d'ordre sémantique et pragmatique à la fois :

[2] 1908-1999, école de linguistique fonctionnelle française.

[3] La dernière relation (3) est plutôt une relation d'équivalence et non pas de dépendance. Elle est pertinente pour des phénomènes comme la coordination ou l'énumération mais elle ne l'est pas pour les phénomènes de clôture.

les parenthèses amènent une information supplémentaire qui par conséquent ne peut pas être l'objet du verbe principal.

L'exemple suivant montre l'existence d'un lien (objet direct) entre un verbe et l'ensemble du segment entre guillemets :

(9) *"Le rapport met en évidence un grave problème", [a déclaré] M. Julian Fantino, chef du service de police de Toronto.*

La relation de type (2) implique la dépendance d'un élément linguistique vers une tête ou noyau (ex. l'adjectif vis à vis du nom duquel il dépend). Nous l'appelons *dépendance indirecte*. Ce type de dépendance concerne généralement les dépendances nominales (la tête est un nom). Pour les phénomènes de clôture, on trouve beaucoup plus fréquemment des parenthèses, mais les guillemets sont également possibles.

Voici quelques exemples de relations entre noms (entourés de crochets) et structures parenthésées, contiguës ou non :

(10) *Près d'un an plus tard pourtant, la chaîne compte toujours le même [nombre] (quatorze) de surfaces de vente.*

(11) *Ce [bouton] (1) libère le mécanisme d'attelage automatique.*

(12) *La [Société] Concessionnaire du Boulevard Périphérique Nord de Lyon (SCBPNL) a été dans l'impossibilité, fin avril, d'arrêter ses comptes 1997.*

Le tableau suivant récapitule les caractéristiques selon le type de dépendance. Il ajoute aussi de l'information concernant la position et la distance de la tête par rapport à l'élément qui en dépend :

	Dépendance Directe	**Dépendance Indirecte**
Nature de la tête	verbe	nom
Nature de la dépend.	OBJ, VADJ, VN, VPP	NN, NADJ, NPP
Phénomène le plus fréquent	« » “ ”	( ) [ ]
Position de la tête	antéposée ou postposée	antéposée
Distance de la tête	contiguë	contiguë ou non contiguë

Finalement, il est a signaler qu'il existe des cas complexes où une information autre que syntaxique (au niveau discursif) serait nécessaire pour retrouver la tête avec laquelle les phénomènes de clôture ont une relation de dépendance. Par exemple :

(13) *Comme ce témoin indique que l'alternateur ne charge pas, vérifier la tension de la courroie d'alternateur/ventilateur (voir page 45).*

(14) *Il a jugé qu'il y avait [TRADUCTION] "atteinte aux droits des appelants".*

(15) *L'immunoscintigraphie (cf annexe J) permettant de mettre en évidence des métastases infracliniques est encore en cours de développement [BLEND1992].*

Pour le premier et pour le troisième exemple, les éléments entre parenthèses et crochets n'ont pas de référent dans la phrase elle-même, mais plutôt au niveau du texte dans lequel ils apparaissent. Dans le deuxième exemple, il s'agit d'une indication pragmatique.

D'après cette caractérisation, nous avons défini l'ensemble de dépendances suivantes pour les phénomènes de clôture[4] :

(1) Dépendances directes

`OBJGM(x,y)`, où `x` est un verbe et `y` une structure entre guillemets objet du verbe principal, par exemple :

Il a notamment refusé de reprendre à son compte la proposition, allant jusqu'à [dire] " Je me refuse à faire croire qu'à partir de mars il y aura de l'argent dans les caisses. "

`VGM(x,y)`, où `x` est un verbe et `y` un attribut ou un prédicatifs entre guillemets, par exemple :

Le gouvernement estime que l'accord [est] « équilibré » car il a reconquis la maîtrise des modalités de privatisation.

(2) Dépendances indirectes

`NGM(x,y)`, où `x` est un nom et `y` un modifieur nominal entre guillemets, par exemple :

Dans une lettre au Premier ministre, le leader syndical met en garde contre de possibles [conséquences] « catastrophiques » pour l'emploi.

Le [thème] "Structure et Dynamique des Atomes et des Ions" a une place toute particulière dans les activités du laboratoire.

`NPR(x,y)`, où `x` est un nom et `y` un modifieur nominal entre parenthèses, par exemple :

L'[hypercapnie] (PaCO2>45mmhg) est un signe de gravité reconnu qui témoigne d'une importante altération du régime ventilatoire.

La surveillance repose sur l'[examen] clinique (standard).

6.3.3 Évaluation du module (ponctuation)

Le corpus utilisé pour évaluer cette grammaire est un corpus d'environ 63.000 mots et 3.000 phrases, extrait du Web (avril 01). Il est constitué de textes de domaines variés

[4]Nous avons mis des crochets autour des têtes syntaxiques de chaque relation en exemple.

(journaux généraux et de finances, documents juridiques et scientifiques, manuel technique).

Sur les 3.000 phrases, nous en avons choisi au hasard 180 avec des guillemets et/ou parenthèses pour évaluer manuellement l'extraction des trois dépendances spécialisées définies dans ce module ainsi qu'une dépendance de base (`SUBJ`) :

Rélation	*Total*	*Extraits*	*Précision*	*Rappel*
NPR	146	102	63 %	70 %
NGM	24	19	100 %	79,2 %
OBJGM	12	9	90 %	75 %
SUBJ	294	251	94,8 %	88,3 %

Les moyennes globales des taux de précision et de rappel pour l'extraction de ces dépendances est de 78,9 % et 80 % respectivement.

6.3.4 Applications possibles

Une approche fonctionnelle pour l'analyse syntaxique de quelques phénomènes liés à la ponctuation permet l'extraction de dépendances syntaxiques entre un élément et l'élément dont il dépend dans une phrase donnée. Ces dépendances peuvent alors être utiles par la suite dans des domaines divers en linguistique comme la lexicographie (enrichissement de dictionnaires), la sémantique (désambiguïsation, ajout d'informations sémantiques), etc. et aussi dans des tâches de question-réponse et/ou de recherche d'information(s).

Traitements linguistiques : lexicographie, sémantique...

* Sigles

  NPR(FEDER,(le Fonds européen de développement régional))
  1225>Il a précisé que , si la région écossaise concernée est en effet éligible au FEDER PR{(le Fonds européen de développement régional)} , le niveau des aides qu' elle apporte à l' opération Hoover est conforme aux règles communautaires .

* Termes, entités (nommées)

  NGM(cultures," à rotation rapide ")
  1612>Elles étaient condamnées à des cultures GM{" à rotation rapide"} .

  NPR(Botswana,(République du))
  441>L' entente actuelle , dont les membres sont l' Afrique du Sud , le Botswana PR{(République du)} , le Lésotho , la Namibie et le Swaziland , a été signée en 1969 .

* Traductions

  NPR(" Restitution of Land Rights Act ",(Loi sur la restitution du droit terrien))
  368>La GM{" Restitution of Land Rights Act"} PR{(Loi sur la restitution du droit terrien)} a été adoptée par le Parlement sud-africain en novembre 1994 .

* Définitions

NPR(Charia,[loi islamique])
1375>" A bas la Charia PR{[loi islamique]} , à bas le Hezbollah .

NPR(démarcheurs,(vendeurs de rue et de marchés aux puces))
298>des démarcheurs PR{(vendeurs de rue et de marchés aux puces)} ;

Autres tâches : systèmes de QA et Recherche d'Information

* Citations

OBJGM(déclaré," Nous avons appris que le crime organisé considère le Canada comme un refuge ")
8274>GM{" Nous avons appris que le crime organisé considère le Canada comme un refuge"} , a déclaré M. Flaherty .

* Titres

NGM(thème," Structure et Dynamique des Atomes et des Ions ")
4119>Le thème GM{" Structure et Dynamique des Atomes et des Ions"} a une place toute particulière dans les activités du laboratoire .

* Équivalents monétaires, pourcentages...

NPR(tournesol,(36 %))
NPR(huile de palme,(25 %))
590>L' huile de tournesol PR{(36 %)} figurait en tête des importations devant l' huile de palme PR{(25 %)} .

NPR(États-Unis,(299 000 $ US))
NPR(Royaume-Uni,(200 000 $ US))
NPR(volaille,(incluant les poules à bouillir , le canard sauvage , les oies , etc. ,))
703>Les plus grands fournisseurs de volaille vivante PR{(incluant les poules à bouillir , le canard sauvage , les oies , etc. ,)} en 1994 étaient les États-Unis PR{(299 000 $ US)} , le Royaume-Uni PR{(200 000 $ US)} et les Pays Bas PR{(67 000 $ US)} .

* Appositions diverses (localisations, dates, surnoms...)

NPR(Ecole,(EHESS , 54 , boulevard Raspail , Paris 6))
1692>L' Ecole des hautes études en sciences sociales PR{(EHESS , 54 , boulevard Raspail , Paris 6)} accueille une exposition (...) .

NPR(L.U.L.I.,(Palaiseau))
NPR(C.E.A.,(Bruyères))
NPR(L.L.N.L.,(Livermore))
4214>Ce formalisme, fait l' objet de collaborations avec des équipes du L.U.L.I. PR{(Palaiseau)} , du C.E.A. PR{(Bruyères)} et du L.L.N.L. PR{(Livermore)} .

* Informations supplémentaires

NPR(Nut,(biscuits))
NPR(Willards,(croustilles de maïs et de pomme de terre))
NPR(Snacks,(maïs éclaté))
875>Bakers et Royal Beech - Nut PR{(biscuits)} , Simba et Willards PR{(croustilles de maïs et de pomme de terre)} et Baker Street Snacks PR{(maïs éclaté)} comptent parmi les plus importants fabricants de ce secteur .

NPR(Bourges,(" le roi des courges "))
2093>La télévision est sa bête noire , de TF 1 à la cérémonie des Césars et Hervé Bourges PR{(" le roi des courges")} (...)

* Précisions (extensions)

NPR(gamme,(Haute , Moyenne et Basse ainsi que la marche arrière , la 1ère , la 2ème et la 3ème vitesses .)
3525>Le levier de changement de vitesses (...) sert à sélectionner la gamme de vitesses du tracteur PR{(Haute , Moyenne et Basse ainsi que la marche arrière , la 1ère , la 2ème et la 3ème vitesses)} .

NPR(Maggie,(Emma Thompson))
NPR(Roger,(Hugh Laurie))
NPR(Mary,(Imelda Staunton))
NPR(Sarah,(Alphonsia Emmanuel))
1269>Maggie PR{(Emma Thompson)} est devenue éditrice , Roger PR{(Hugh Laurie)} et Mary PR{(Imelda Staunton)} se sont mariés et écrivent des jingles publicitaires , Sarah PR{(Alphonsia Emmanuel)} est restée actrice et collectionne les amants .

6.4 Traitement des titres

Le traitement des titres en tant qu'unités linguistiques ou textuelles se justifie principalement par rapport à leur structure : les titres se différencient des autres unités textuelles du fait d'être physiquement séparés des unités suivantes par des blancs verticaux (généralement deux retours à la ligne). À la différence d'autres phrases, ils peuvent avoir des blancs horizontaux au début et à la fin (cas de titres centrés par rapport au corps du texte).

Le traitement précis des titres par le parseur peut avoir des applications notamment dans le domaine de la recherche et l'extraction d'information (extraction de documents, indexation, représentation...) [HZ80], [Kwo75], [LM98].

6.4.1 Segments spécialisés

Comme nous l'avons présenté au chapitre précédent (*cf.* 5.2) avant notre étape de marquage en segments (*chunking*), il existe une étape préalable où nous reconnaissons des phrases comme étant des titres (avec un taux de succès de reconnaissance de 96,6 % .

Exemple pour le titre suivant, après balisage :

(16) *1. 3 Rappel succinct sur les quasi-particules*

```
<tit>  1 <sp>.</sp> 3 Rappel succinct sur les quasi-particules
 </tit>
```

En ce qui concerne le marquage de segments proprement dits, dans le cas de titres, il n'existe pas de segment spécialisé comme c'est le cas pour d'autres phénomènes spécialisés. Le *chunking* des titres se résume en deux étapes : le marquage éventuel de phénomènes spécialisés (concernant d'autres phénomènes comme les séparateurs, les organisateurs, les guillements, les parenthèses) et le marquage des syntagmes de base.

Voici le résultat pour le titre de l'exemple précédent :

```
57>MAX{<tit> SP{1 <sp>.</sp>} SP{3} NP{Rappel} AP{succinct}
PP{sur NP{les NOUN{quasi- particules}}} </tit>}
```

Dans cet exemple, le texte en entrée pour l'étape d'extraction de dépendances est un texte balisé (avec des marques de type `<tit>`) et segmenté en syntagmes noyaux (`NP`, `FV`, etc.).

6.4.2 Dépendances

En ce qui concerne les fonctions syntaxiques à l'intérieur des titres[5] nous considérons que, en plus des différentes relations de dépendance établies entre les têtes des constituants, il nous semble important de créer une relation unaire concernant l'élément clef du titre (le nom tête du premier `NP`).

La tête de la relation `CLE( )` est toujours un nom (ou deux s'ils sont coordonnés) ; ce n'est jamais un chiffre ou un organisateur. Nous ne tenons pas compte non plus d'autres catégories (verbes, par exemple) car d'après notre étude de corpus les titres sont majoritairement composées de structures nominales. Si le titre est constitué d'une série de `NP` avec des extensions variées ou bien des propositions entières (avec verbe conjugué), le nom extrait est alors le sujet de la proposition.

Voici deux exemples :

(17) *LES FREINS*

```
CLE(FREINS)
13>MAX{<tit> NP{LES FREINS} </tit>}
```

(18) *Cas où l'immeuble est acquis par l'EURL ou porté à l'actif du bilan de l'exploitation individuelle*

[5] Les fonctions des titres par rapport au texte vont au-delà de la syntaxe et donc des limites de notre analyseur.

```
CLE_57(Cas)
86>MAX{<tit> NP{Cas} SBC{BG{où} NP{l' immeuble} FV{est acquis}}
PP{par NP{l' EURL}} ou AP{porté} PP{à NP{l' actif}} PP{du
NP{bilan}} PP{de NP{l' exploitation}} AP{individuelle} </tit>}
```

La dépendance `CLE( )` peut être extraite ou déduite dans le cas d'énumérations et coordinations. La règle d'extraction de la relation `CLE( )` est la suivante :

```
CLE(#1) = punct[form:fdebtit,sgml:+,start:+,first:+],
SP*, NP{?*, #1[last:+,pron:~,num:~]}}.
```

La clé d'un titre est le premier nom après la balise `<tit>` et éventuellement quelques séparateurs. La balise `<tit>` est une ponctuation qui a comme traits la forme `fdebtit` ainsi que les traits `sgml`, `start` et `first` en positif. Les traits `start` et `first` correspondent à des marques de début d'arbre et de *chunk* respectivement.

La règle de déduction pour l'extraction d'une clé dans le cas d'une coordination est :

```
| NP{?*, #1[last:+,pron:~,num:~]}, PP*, AP*,
coord, NP{?*, #2[last:+,pron:~,num:~]} |

if ( cle(#1[noun:+]))
cle(#2).
```

Cette règle nous indique que la clé d'un titre peut être aussi la tête du deuxième `NP` si celui est coordonné et si le nom du premier a déjà été extrait comme clé. Par exemple :

(19) *3.3- L'évolution de la maîtrise et l'organisation du travail*

```
CLE(évolution)
CLE(organisation)
NPP(évolution,de,maîtrise)
NPP(organisation,du,travail)
632>MAX{<tit> SP{3 <sp>.</sp>} SP{3 -} NP{L' évolution} PP{de
NP{la maîtrise}} et NP{l' organisation} PP{du NP{travail}} </tit>}
```

(20) *Collecte et conservation des informations nominatives*

```
CLE(Collecte)
CLE(conservation)
NPP(conservation,des,informations)
6>MAX{<tit> NP{Collecte} et NP{conservation}
PP{des NP{informations}} AP{nominatives} </tit>}
```

À notre connaissance, il n'existe pas d'autres systèmes d'analyse syntaxique prenant en compte de manière spécifique les titres.

6.4.3 Évaluation du module (titres)

Le corpus utilisé pour l'évaluation de cette grammaire est un corpus de domaines variés constitué de 55.103 mots, avec une moyenne de 18 mots par phrase de type N1 (environ 24 % du corpus) et de 38 mots par phrase de type N2 (environ 76 % du total des phrases du corpus).

Parmi ces 76 % de phrases de deuxième niveau il y a 13,3 % de titres. Le taux de précision du balisage est de 95,2 % (sur 266 titres marqués, 253 sont corrects). Le tableau suivant montre les résultats des taux de précision et de rappel pour les dépendances `CLE( )` et `SUBJ( )` :

	Précision	*Rappel*	*F1*
CLE( )	93 %	89,6 %	91,3 %
SUBJ( )	100 %	100 %	100 %

La dépendance `SUBJ` est extraite dans les cas de titres avec verbe conjugué (20 % des titres de l'ensemble du corpus, plus fréquents dans les corpus journalistiques, rares dans les corpus techniques).

Les résultats sont satisfaisants car la dépendance spécialisée présente un taux de précision > 90 % sans pour autant nuire à l'extraction de la dépendance de base (sujet).

6.5 Traitement des listes

Si l'on tient compte de la cohésion des éléments d'une liste, il semble nécessaire de traiter cette structure en tant qu'unité linguistique au même titre que la phrase. Pour ce faire, nous avons vu précédemment au chapitre 5 (*cf.* 5.2.2) que l'organisation des éléments qui l'intègrent requiert la mise en place d'un prétraitement particulier (balisage) pour la délimitation de ses frontières. Cela étant fait, l'analyse syntaxique des listes se fait comme pour les autres phrases présentant des phénomènes de deuxième niveau.

Concrètement, l'analyse syntaxique des listes implique un traitement de la ponctuation et surtout des aspects liés à la structure du document (espaces horizontaux et verticaux). L'extraction de dépendances est aussi adaptée à ce phénomène et, comme nous allons le montrer, elle met en avant l'existence de relations sémantiques qui pourront être utilisées par la suite dans de nombreuses applications (construction de dictionnaires, désambiguïsation sémantique, etc.).

6.5.1 Segments spécialisés

Pour rappel, une *liste* est une structure composée d'une amorce et de plusieurs items (au moins deux) disposés verticalement. Un *séparateur* est un signe de ponctuation qui délimite deux éléments.

Comme nous l'avons vu, le prétraitement des listes consiste en l'introduction des balises <list> et < /list> en début et fin de liste et de <sp> et < /sp> en début et fin de séparateur. Dans les listes, nous avons tenu compte des cinq séparateurs suivants :

<sp> : < /sp> <sp> ? < /sp>

<sp> ; < /sp> <sp> ! < /sp>

<sp> . < /sp>

Tous ces symboles de ponctuation entourés de balises ont été ajoutés au dictionnaire morphologique pour qu'ils soient reconnus en tant que *tokens* par le parseur. De même, un nouveau symbole de fin de phrase a été inclus dans le segmenteur : en plus de l'étiquette +SENT (*sentence*) nous avons introduit la marque < /list>. Quand notre analyseur trouve l'une de ces deux étiquettes, il marque la fin d'une phrase.

L'étape de marquage en constituants (*chunking*) se résume en trois étapes :

a) marquage de séparateurs et/ou organisateurs (`SP{ }`) ;

b) marquage de syntagmes de base et d'autres segments spécialisés (le cas écheant) ;

c) marquage d'items (`IT{ }`) et amorces (`AMR{ }`).

Voici un exemple de liste en entrée :

(21) *" Les traitements ont été regroupés en 2 catégories :*
- courte durée et fort débit, de quelques secondes à quelques minutes ;
- longue durée et débit plus faible, de 30 mn à quelques heures. "

Après le prétraitement (balisage), ce même exemple devient :

```
<list> Les traitements ont été regroupés en 2 catégories <sp>:</sp>
- courte durée et fort débit, de quelques secondes à quelques minutes
<sp>;</sp>
- longue durée et débit plus faible, de 30 mn à quelques heures
<sp>.</sp> </list>
```

Résultat après pré-traitement et analyse par le chunker :

```
8>MAX{<list> AMR{NP{Les traitements} FV{ont été regroupés} PP{en
NP{2 catégories}} <sp>:</sp>} IT{SP{-} NP{AP{courte} durée} et
NP{AP{fort} débit} , PP{de NP{quelques secondes}} PP{à NP{quelques
minutes}} <sp>;</sp>} IT{SP{-} NP{AP{longue} durée} et NP{débit}
AP{plus faible} , NP{de NOUN{30 mn}} PP{à NP{quelques heures}}
<sp>.</sp>} </list>}
```

Nous décrivons par la suite la nature et les caractéristiques des segments spécialisés que nous avons introduits et qui sont particuliers aux listes.

Séparateurs et organisateurs

Le segment SP{} délimite un séparateur ou un organisateur en début d'item. Pour le marquage de ce type de chunk on utilise une quinzaine de règles dites de séquence (*cf.* 4.5.1).

Les cas les plus simples à traiter concernent le marquage des séparateurs ([*], [•], [-]) ainsi que des chiffres suivis d'une parenthèse fermante (*1)*) ou d'un tiret (*1-*) (à condition qu'il n'y ait pas de parenthèse ouvrante ou de tiret avant). L'exemple[6] suivant montre des séparateurs «simples» :

```
SP{-} NP{Admission} AP{définitive} (...) <sp>.</sp>
SP{-} NP{Liste} PP{d' NP{attente}} (...) <sp>.</sp>
SP{-} NP{Avis} PP{de NP{refus}} (...) <sp>.</sp>
```

Le cas des organisateurs comportant des lettres suivies soit d'un point (*a.*) soit d'une parenthèse fermante (*a)*) s'avère plus complexe à traiter, car les lettres sont souvent étiquetées en tant que noms (il y a alors un risque d'ambiguïté avec d'autres noms suivis de points, par exemple des items composés d'un seul nom).

Voici un exemple pour des séparateurs plus « complexes » :

```
SP{a <sp>.</sp>} NP{l' interdiction} (...) <sp>;</sp>
SP{b <sp>.</sp>} NP{l' obligation} (...) <sp>;</sp>
SP{c <sp>.</sp>} NP{le consentement} (...) <sp>.</sp>
```

Le marquage de séparateurs facilite le marquage ultérieur des items des listes.

Amorces des listes

Pour le marquage des amorces, nous avons créé une nouvelle structure (AMR{}) qui délimite la séquence introductrice des listes. On utilise quatre règles de séquence qui s'appliquent une fois que le découpage en syntagmes de base est fait.

En général, la limite gauche de l'AMR est un début de phrase et la limite droite deux points (exemple 22). Si lors du marquage de la liste, plus d'une phrase a été incluse avant l'énumération des items, seulement la dernière sera considérée amorce (exemple 23).

Exemples :

(22) *" L'unité de succion PRS-FIX comprend deux parties principales : "*

```
35>MAX{<list> AMR{NP{L' unité} PP{de NP{succion}} NP{NOUN{PRS - FIX}}
FV{comprend} NP{deux parties} AP{principales} <sp>:</sp>}
```

(23) *" En procédant de la sorte, nous avons pu mettre en évidence les ingrédients nécessaires à la production d'un catalyseur pour la réduction de l'oxygène en PEFCs. Ces ingrédients sont : "*

[6] Le contenu des items de la liste en exemple est volontairement simplifié pour une meilleure visualisation de la sortie du parseur à cet stade.

```
46>MAX{<list> GV{En procédant} de la sorte , NP{nous} FV{avons pu}
IV{mettre} PP{en NP{évidence}} NP{les ingrédients} AP{nécessaires} PP{à
NP{la production}} PP{d' NP{un catalyseur}} PP{pour NP{la réduction}}
PP{de NP{l' oxygène}} PP{en NP{PEFCs}} <sp>.</sp> AMR{NP{Ces ingrédients}
FV{sont} <sp>:</sp>}
```

Une fois que l'amorce de la liste est délimitée, le parseur identifie les différents items intégrant la liste.

Items

Le segment `IT{}` délimite les items des listes. Dix règles de séquence permettent le marquage de ces unités.

Comme pour les amorces (`AMR{}`) le noyau (tête syntaxique) n'est pas forcément l'élément le plus à droite. Au contraire, le noyau est la tête du premier élément de l'item ou la tête du deuxième s'il n'y a pas de chunk `SP{}` (les têtes sont extraites plus tard dans les dépendances).

En général, les items ont comme premier élément un segment `SP{}` et une balise `<sp>.</sp>` à la fin.

Exemple :

(24) *" - Faible voltage de circuit,*
- Faible résistance,
- Situation à l'intérieur du compartiment moteur,
- Non écrantement par du métal. "

```
IT{SP{-} NP{AP{Faible} voltage} PP{de NP{circuit}} <sp>,</sp>} IT{SP{-}
NP{AP{Faible} résistance} <sp>,</sp>} IT{SP{-} NP{Situation} PP{à
l'intérieur du NP{compartiment}} AP{moteur} <sp>,</sp>} IT{SP{-} Non
écrantement PP{par NP{du métal}} <sp>.</sp>} </list>}
```

Le résultat du chunking pour la liste ci-dessous est le suivant :

(25) *Pour ce faire, plusieurs techniques sont possibles :*
- le blindage (blindage de matériaux) ;
- le filtrage de type VETO par détecteurs en "sandwich" ;
- l'intégration temporelle des résultats par filtrages numériques, tant spatial que temporel.

```
0>MAX{<list> AMR{Pour ce faire , NP{plusieurs techniques} FV{sont}
AP{possibles} <sp>:</sp>} IT{SP{-} NP{le blindage} PR{( NP{blindage}
PP{de NP{matériaux}} )} <sp>;</sp>} IT{SP{-} NP{le filtrage} PP{de
NP{type}} NP{VETO} PP{par NP{détecteurs}} PP{en GM{" NP{sandwich} "}}
<sp>;</sp>} IT{SP{-} NP{l' intégration} AP{temporelle} PP{des
NP{résultats}} PP{par NP{filtrages}} AP{numériques} , AP{tant spatial}
que AP{temporel} <sp>.</sp>} </list>}
```

6.5.2 Dépendances

Dans le cas des listes, nous calculons par extraction ou par déduction tout d'abord les noyaux des items, ensuite les noyaux des amorces et les dépendances syntaxiques entre un noyau et les différents items de la liste. Finalement, on extrait le lien sémantique entre l'amorce et un item.

Pour les items et les amorces, quand nous parlons de « noyau » nous faisons référence au « noyau sémantique ». Pour les items et les amorces, donc, l'analyseur n'extrait pas le « noyau syntaxique » (cas des *chunks*) mais le « noyau sémantique ».

Items des listes

La dépendance `ITMLIST( )` extrait les noyaux des items des listes. Comme pour les noyaux des amorces, il s'agit d'une dépendance unaire[7], ou plus précisément d'une sorte de trait donné aux items des listes (nous avons défini –plus loin dans cette même section– une dépendance qui lie chaque item à sa tête syntaxique, en l'occurrence avec une relation d'objet).

En général, l'item est toujours la tête du premier `NP` après un séparateur (si ce n'est pas un pronom). Dans les cas de coordination, on extrait les deux têtes des `NP` coordonnés. Exemples :

```
ITMLIST(caractéristiques)
ITMLIST(services)
ITMLIST(personne)
```

(26) *“ La demande d'avis ou la déclaration doit préciser :*
la personne qui présente la demande (...) ;
les caractéristiques du traitement ;
les services chargés de mettre en oeuvre celui-ci ”

Seulement dans deux cas précis, on a considéré plus informatif que la dépendance `ITMLIST` présente trois éléments. C'est le cas des négations et le cas de pronoms démonstratifs[8].

Exemples :

```
ITMLIST(Arrêt)
ITMLIST(Pas,d',arrêt)
```

(27) *“ Ainsi, outre les pannes de phares, et selon les deux montages possibles, on pourrait assister à deux sous-cas typiques possibles sur les Diesel :*
- Arrêt moteur (blocage de l'électrovalve fonctionnant en émission),
- Pas d'arrêt moteur (électrovalve fonctionnant en manque). ”

[7] On parle de dépendance dans le sens où l'item *dépend* syntaxiquement de l'amorce. À la différence d'autres dépendances, ici l'élément « gouverneur » n'est pas affiché car il s'agit de l'amorce entière.

[8] Par rapport à la liste, le “gouverneur” de la dépendance est toujours l'amorce -qu'on n'affiche pas.

```
ITMLIST(celle,des,pays)
ITMLIST(celle,des,entreprises)
```

(28) *" Parallèlement, les entreprises se trouvent confrontées à une forte concurrence, généralement de deux natures différentes :*
- celle des pays du tiers monde aux très faibles coûts (...)
- celle des autres entreprises aux techniques de production (...) "

Un cas problématique est celui des items qui sont des phrases entières. Exemple[9] :

(29) *" Cette analyse permet de définir de manière pertinente des principes de conception concernant l'outil informatique :*
1- Il doit être considéré comme un élément d'un "système d'aide" à l'action collective.
2- Il doit faciliter la recherche d'informations techniques.
3- Son développement doit s'accompagner d'une gestion de l'information. "

```
227>MAX{<list> AMR{NP{Cette analyse} FV{permet} IV{de définir}
PP{de NP{manière}} AP{pertinente} PP{des NP{principes}} PP{de
NP{conception}} PP{concernant NP{l' outil}} AP{informatique}
<sp>:</sp>}

IT{SP{1 -} NP{Il} FV{doit} IV{être considéré} comme NP{un élément}
PP{d'un GM{" NP{système}} NP{d'aide} "} PP{à NP{l'action}}
AP{collective} <sp>.</sp>}

IT{SP{2 -} NP{Il} FV{doit} IV{faciliter} NP{la recherche} PP{d'
NP{informations}} AP{techniques} <sp>.</sp>}

IT{SP{3 -} NP{Son développement} FV{doit} IV{s' accompagner} PP{d'
NP{une gestion}} PP{de NP{l' information}} <sp>.</sp>} </list>}
```

L'analyseur extrait dans ces cas la tête du premier syntagme nominal (s'il n'est pas un pronom).

Noyaux des amorces

Le noyau d'une amorce, `NOYAU(x)`, est son `NP` principal (du point de vue sémantique)[10]. Il est calculé moyennant des règles de déduction, une fois que les fonctions syntaxiques de base ont été extraites.

Nous avons identifié différents cas de figure, entre autres :

[9] Les coupures entre l'amorce et le premier item et entre les items sont faites pour une meilleure visualisation ; la sortie du parseur est une structure plus compacte.

[10] Nous n'avons pas modélisé les cas où l'amorce ne contient pas de `NP` (autre que pronom), par exemple dans des instructions explicites (*« Pour l'utiliser correctement, vous devez installer : le système d'exploitation, Windows 2000, le Navigateur, (...) »*).

(A) Si l'amorce présente un `NP` sujet et un `NP` objet, le noyau est le `NP` objet, exemple :

`NOYAU(critiques)`

"On peut avancer les critiques suivantes, parmi beaucoup d'autres :
- la divergence élevée des résultats (...)
- la définition du protocole d'expérience (...) "

(B) Si le verbe fini de l'amorce a un argument régi, le noyau est le `NP` de cet argument, exemple :

`NOYAU(critères)`

" L'évaluation des conséquences économiques de la non-qualité repose principalement sur deux critères :
1. Le coût de l'obtention de la qualité (...)
2. Les pertes de recettes (...) "

(C) Si l'amorce a comme verbe principal le verbe être et s'il n'y a pas de `NP`s en complément, le noyau est le `NP` sujet, par exemple :

`NOYAU(organismes)`

"Les organismes représentant le secteur alimentaire en Afrique du Sud sont :
la South African Association for Food Science and Technology
la Grocery Manufacturers Association "

(D) Si l'amorce ne contient pas de `NP` après le verbe ni de `PP` en argument, le noyau est le `NP` sujet, par exemple :

`NOYAU(actions)`

" Ainsi, les actions de formation mises en place se caractérisent comme suit :
- formation qualifiante pour les opérateurs ;
- familiarisation aux technologies nouvelles "

(E) Si l'amorce n'a pas de verbe fini, le noyau est le premier `NP`, par exemple :

`NOYAU(Température)`

" Température et temps minimum à respecter :
- 180° pendant 30 min
- 170° pendant 1h00 "

(F) Si l'amorce contient un syntagme nominal (dont le nom est déjà extrait comme possible amorce) suivi d'un syntagme prépositionnel (préposition *de*), on extrait aussi comme noyau le nom qui suit la préposition, par exemple :

`NOYAU(pièces)`

" Il doit être possible de déconnecter l'alimentation des pièces suivantes :
1. Imprimante
2. GEM
3. Succion PRS-FIX "

Nos règles de déduction permettent aussi l'élagage de faux noyaux dans des cas précis. Ainsi, quand deux noyaux sont en concurrence, nous éliminons toujours celui qui est plus près du début de l'amorce (et nous gardons celui qui est plus près de la fin). Dans les cas des `NP` suivis de `PP`, nous gardons les deux noyaux extraits car il n'est pas toujours possible de décider lequel des deux noms est le noyau[11].

Ainsi, si dans l'exemple précédant *« pièces »* est clairement le noyau alors qu' *« alimentation »* ne l'est pas, dans l'exemple suivant on considère que c'est plutôt l'inverse (*« catégories »* est le noyau et non pas *« orbites »*)[12] :

(30) *" On distingue trois catégories d'orbites :*
- LEO (Low Earth Orbit).
- MEO (Medium Earth Orbit).
- GEO (Geostationnary Earth Orbit). "

Autres dépendances

Dans le cas d'amorces non saturées (ANS) (*cf.* 3.4.2), les différents items de la liste remplissent des fonctions syntaxiques par rapport au verbe principal de l'amorce. Les items peuvent ainsi être des sujets, des objets, des arguments (via préposition), etc.

Dans le cas d'amorces saturées (AS) ces relations de dépendance n'existent pas car l'amorce est une séquence syntaxiquement complète.

Nous renvoyons le lecteur à l'Annexe D.5 pour une description détaillée de l'ensemble de ces dépendances avec leurs exemples (`SUBJLIST()`, `VARGLIST()`, `COMPLIST()`, `VNLIST()`, `VADJLIST()`).

Liens sémantiques

Comme [Hea92] ou [FLM99], nous nous sommes intéressée à l'extraction automatique de relations sémantiques.

Dans le cadre des listes, notre analyseur calcule le lien sémantique entre le noyau de l'amorce et les items et produit la relation `RELSEM(x,y)` qui exprime le fait que `x` et `y` sont liés sémantiquement (`x` et `y` sont des noms). Ce lien est uniquement produit dans les cas d'amorces saturées (AS). Il s'agit d'une *équivalence* syntaxique et non pas d'une réelle *dépendance*.

[11] On perd de la précision mais on gagne en rappel (*cf.* 6.5.3)

[12] On aurait pu considérer aussi *« une catégorie de »* comme une sorte de déterminant nominal et garder *« orbites »* comme le vrai noyau.

Il existe un lien sémantique similaire entre les éléments d'une relation `NPR` (phénomènes de clôture), c'est-à-dire entre un nom et le contenu de la parenthèse (s'il s'agit d'un ou plusieurs noms énumérés). Exemple :

(31) *"Certaines personnes sont plus vulnérables à la pollution (...) en particulier vis-à-vis des polluants oxydants (ozone, oxyde d'azote)."*

```
NPR(polluants, (ozone, oxyde d'azote))
```

Les relations dans les listes et dans les parenthèses contenant des énumérations que nous avons modélisées sont de type hyperonymie-hyponymie (*est-un*) ou holonymie-méronymie (*fait-partie-de*).

Pour les listes, la stratégie adoptée par notre analyseur est prudente dans le sens qu'il se limite à montrer l'existence d'un lien sémantique (essayer d'établir la différence entre hyperonymie-hyponymie et holonymie-méronymie avec des moyens uniquement morpho-syntaxiques implique un risque élevé d'erreurs) :

(32) *" Il est actuellement possible de corriger les affections suivantes :*
- Myopie .
- Hypermétropie .
- Astigmatisme . "

```
RELSEM_[28](affections,Myopie)
RELSEM_[28](affections,Hypermétropie)
RELSEM_[28](affections,Astigmatisme)

338>MAX{<list> AMR{NP{Il} FV{est} AP{actuellement possible} IV{de corriger}
NP{les affections} AP{suivantes} <sp>:</sp>} IT{SP{-} NP{Myopie} <sp>.</sp>}
IT{SP{-} NP{Hypermétropie} <sp>.</sp>} IT{SP{-} NP{Astigmatisme} <sp>.</sp>}
</list>}
```

Résultat final de l'analyse

Après la segmentation et l'extraction des dépendances, pour la liste de l'exemple (25) nous obtenons la sortie suivante[13] :

```
NOYAU_[8](techniques)
ITMLIST_35(blindage)
ITMLIST_35(filtrage)
ITMLIST_35(intégration)
RELSEM_[28](techniques,blindage)
```

[13] L'option d'affichage montre le numéro de règle utilisée. Le chiffre entre crochets indique qu'il s'agit d'une règle de déduction, sinon c'est une règle d'extraction.

```
RELSEM_[28](techniques,filtrage)
RELSEM_[28](techniques,intégration)

0>MAX{<list> AMR{Pour ce faire , NP{plusieurs techniques} FV{sont}
AP{possibles} <sp>:</sp>} IT{SP{-} NP{le blindage} PR{( NP{blindage}
PP{de NP{matériaux}} )} <sp>;</sp>} IT{SP{-} NP{le filtrage} PP{de
NP{type}} NP{VETO} PP{par NP{détecteurs}} PP{en GM{" NP{sandwich} "}}
<sp>;</sp>} IT{SP{-} NP{l' intégration} AP{temporelle} PP{des
NP{résultats}} PP{par NP{filtrages}} AP{numériques} , AP{tant spatial}
que AP{temporel} <sp>.</sp>} </list>}
```

Le module d'extraction de dépendances de base et éventuellement les autres grammaires spécialisées ajoutent les dépendances correspondantes, en l'occurrence :

```
NGM_34(détecteurs," sandwich ")
NPR_14(blindage,( blindage de matériaux ))

SUBJ_61(sont,techniques)
NN_86(type,VETO)
NADJ_91(intégration,temporelle)
NADJ_91(filtrages,numériques)
NADJ_92(résultats,numériques)
VADJ_67(sont,possibles)
```

6.5.3 Évaluation du module (listes)

Pour évaluer les résultats obtenus par la grammaire des listes nous avons utilisé un nouveau corpus d'environ 35.000 mots que nous caractérisons plus loin. Après une première étape d'analyse et de tri des phrases par la grammaire noyau, les phrases N2 ont été pré-traitées (balisage pour listes) et après analysées par les grammaires spécialisées (phénomènes de clôture et phénomènes de séparation). L'évaluation a été faite manuellement.

En ce qui concerne le *chunking*, nous avons évalué le marquage correct des structures nouvelles, `AMR` et `IT`. Pour les dépendances, nous avons évalué l'extraction des noyaux sémantiques des amorces et des items, l'extraction de dépendances syntaxiques entre amorces et items et l'extraction de liens sémantiques. Finalement, nous avons évalué l'extraction de deux dépendances de base : `SUBJ` (sujet) et `COMP` (complément d'objet direct).

Corpus d'évaluation

Le corpus d'évaluation pour ce module est un nouveau corpus récupéré sur le Web (août 01). Il est constitué de textes de domaines variés (juridique, scientifique-technique, économique). Le tableau suivant montre certaines de ses caractéristiques :

Corpus	*Nombre de mots*	*Nombre de phrases*	*Nombre de listes*
juridique	13.065	620	33
scientifique-technique	9.924	533	23
économique	12.203	527	20
total	35.192	1.680	76

Les caractéristiques générales des 76 listes sont les suivantes : les amorces contiennent 1.452 mots, ce qui représente une moyenne de 19 mots par amorce. Il y a un total de 246 items (5.342 mots) avec une moyenne de 21 mots par item. Les listes représentent 406 phrases, c'est à dire 24,2 % des phrases du total du corpus.

Le tableau suivant présente la distribution de toutes les phrases des corpus après analyse par la grammaire noyau et par les deux grammaires spécialisées existantes (traitement de phénomènes de clôture et de phénomènes de séparation) :

Type de phrases	*Nombre de mots*	*Nombre de phrases*	*Pourcentage de phrases*
Phrases N1			
total	6.362	350	**20,8 %**
Phrases N2			
listes	6.794	406	**24,2 %**
total N1 et N2	35.192	1.680	**100 %**

La moyenne de mots/phrase est de 18 pour les phrases N1 et de 22 pour les phrases N2. Par type de corpus, les phrases N2 ont 19 mots/phrase dans les corpus scientifique-technique, 22 mots/phrase dans le corpus juridique et 25 mots/phrase dans le corpus économique. La moyenne pour les phrases N1 reste 18 mots pour tous les types de corpus.

Résultats du chunking spécialisé

Le tableau suivant présente les résultats de l'évaluation pour le marquage d'amorces et items : précision (`P`), rappel (`R`) et `F1`[14] :

	Précision	*Rappel*	*F1*
`AMR{}`	100 %	98,7 %	99,3 %
`IT{}`	83,4 %	100 %	90,9 %

Il y a une seule `AMR{}` qui n'a pas été marquée (sur un total de 76). Il s'agit d'un cas de variation typographique rare (fin avec un point et non pas deux points) qui n'est pas pris en compte par les règles de la grammaire.

Pour les items, parmi les 295 items marqués, 246 sont corrects et 49 sont faux. Les structures erronées correspondent à des items longs composés de plusieurs lignes.

[14] F1 = 2(P*R)/P+R

Résultats de l'extraction de dépendances spécialisées

Le tableau suivant montre les résultats concernant l'extraction et la déduction de dépendances spécialisées des listes :

	Précision	*Rappel*	*F1*
NOYAU	76,8 %	81,5 %	79,1 %
ITMLIST	90,4 %	97,7 %	93,9 %
RELSEM	69,6 %	87,2 %	77,4 %

Pour la relation NOYAU[15], parmi les 11 noyaux existants qui n'ont pas été extraits, certains se trouvent dans des syntagmes prépositionnels (*"orienter les travaux dans deux directions principales", "aura pour conséquences", "par l'intermédiaire des relations suivantes"*), d'autres ont été mal étiquetés.

La plupart des 23 ITMLIST extraits erronés correspondent à des NPs (subj) dans des items qui sont des phrases complètes. Le rappel (5 items non identifiés) correspond principalement à des erreurs de tagging.

La relation RELSEM dépend totalement de l'extraction/déduction des noyaux d'amorces et d'items. Il y a donc des erreurs quand les relations NOYAU et ITMLIST sont erronées, par exemple dans le cas où l'item est une phrase entière.

Dans certaines listes, la relation RELSEM n'est pas très informative (le noyau est un nom assez générique : *"axes", "questions"*) mais dans d'autres elle est plus significative : *"thèmes, société/publicité/immobilier", "caractéristiques, débit/puissance", "paramètres, distance/diamètre/fréquence", etc.*.

Résultats de l'extraction de dépendances de base

Le tableau suivant montre les résultats pour les sujets et compléments d'objet direct. On a évalué les 76 listes ainsi qu'une centaine de phrases N2 choisies au hasard parmi celles traitées par le module dédié aux phénomènes de clôture :

	Précision	*Rappel*	*F1*
SUBJ	90,4 %	92,7 %	91,5 %
COMP	91,2 %	97,2 %	94,1 %

La principale cause d'erreur pour les deux relations est l'incorrection de l'étiquetage, particulièrement pour des noms étiquetés comme verbes (*« l'arnaque », « le solde », « la date », « le but »* etc.), mais aussi des verbes étiquetés comme noms (*« réduit »*) ou adjectifs (*« porte »*). (Il est à noter que certaines parties du corpus en entrée présentent des erreurs de découpage/écriture de mots.)

Même avec ces erreurs, les taux de précision et de rappel sont supérieurs à 90 %.

[15] Il y a moins de noyaux que de listes (65 noyaux pour 76 listes) car on considère que certaines listes n'en ont pas (i.e. quand les items sont des verbes à l'infinitif, quand l'amorce contient un pronom anaphorique).

6.6 Traitement des énumérations

Comme nous l'avons vu au troisième chapitre, les énumérations constituent un cas spécifique de liste : les différents items qui les intègrent sont équivalents du point de vue de leur fonction syntaxique mais ils ne présentent pas de structure visuelle particulière (ils apparaîsssent les uns après les autres séparés par des symboles de ponctuation).

Pour le traitement spécialisé des énumérations, il s'avère nécessaire de prendre en compte des éléments liés au traitement de la ponctuation en ce qui concerne le marquage structurel et des liens sémantiques similaires aux listes en ce qui concerne l'extraction de dépendances.

6.6.1 Segments spécialisés

Le marquage de segments spécialisés se fait après le repérage de syntagmes de base. La structure `ENM{ }` délimite une énumération. Comme pour les listes et les phénomènes de clôture, il s'agit d'une macro-structure qui contient plusieurs syntagmes noyau à l'intérieur.

Nous avons considéré deux types d'énumération selon le nombre d'items. Le premier type est constitué d'énumérations à deux items. Trois règles de grammaire ont suffi pour les caractériser :

- deux `NP` coordonnés après la conjonction *comme* ;
- deux `NP` coordonnés après l'adverbe *notamment* ;
- deux `NP` coordonnés ou séparés par une virgule à l'intérieur de parenthèses.

Exemples :

(33) *"Certains se déprécient régulièrement par l'obsolescence technologique , [comme les bâtiments et les outillages] et donnent lieu à un amortissement régulier."*

(34) *"Fruits et fruits à coque [(noix , amandes ...)] "*

Le deuxième type d'énumération concerne les énumérations présentant trois items ou plus (ces énumérations sont définies par une douzaine de règles de grammaire). La liste de cas modélisés est restreinte pour éviter le marquage de suites de `NP` qui ne sont pas des vraies énumérations.

Voici quelques cas de ce que nous avons défini comme énumérations :

- plusieurs `NP`, les deux derniers éventuellement coordonnés, après la conjonction *comme* ou après les adverbes *notamment*, *spécialement*, *particulièrement*[16] ;
- plusieurs `NP`, les deux derniers éventuellement coordonnés, finissant par l'adverbe *etc.* ou par la ponctuation ... ;
- une suite de `NP` à l'intérieur de parenthèses, crochets ou tirets ;
- une suite de `NP` en fin de phrase (ponctuation forte à la fin) ;
- une suite de `PP` en fin de phrase (ponctuation forte à la fin).

Exemples :

[16] Ces adverbes ont été choisis comme débuts d'énumération après une étude sur corpus.

(35) *"Les charges (...) comprennent notamment [les frais de personnel , les achats , la sous-traitance , les frais financiers , les charges exceptionnelles et les impôts] .*

(36) *"En choisissant un nombre restreint d' états atomiques discrets résonnants (...) il est possible de définir des grandeurs atomiques [- largeurs d' ionisation , déplacements lumineux , fréquences de Rabi , couplage Raman ... -] représentant les interactions de rayonnement (...) ."*

(37) *"La nouveauté n' est -elle pas la percée de la gauche dans des arrondissements plus centraux comme [le IXe , le IIe , le IIIe ou le IVe ?] "*

Trois règles dites de « précédence linéaire » (*cf.* 4.5.1) permettent d'affiner l'ensemble des règles de marquage des énumérations. En effet, ces règles contrôlent la position de certains éléments à l'intérieur de l'énumération (il ne s'agit pas d'une précédence immédiate). Ainsi, une virgule doit apparaître avant une coordination, les points de suspension doivent apparaître après une virgule ou après un constituant avec le trait nom.

```
8>[form:fcm] < [coord:+]
8>[form:fcm] < punct[form:f3p]
8>[noun:+]  < punct[form:f3p]
```

Il existe des cas complexes du point de vue du marquage : énumérations imbriquées où une énumération en intègre une autre. Dans ces cas, seule l'énumération « intérieure » est repérée :

(38) *"Il y a certains aliments demandés par l' industrie sud-africaine qui ne sont pas produits au pays , notamment [les figues séchées , des fruits comme [les cerises noires et les baies] , et les haricots secs .] "*

```
21>MAX{NP{Il} FV{y a} NP{AP{certains} aliments} AP{demandés}
PP{par NP{l' industrie}} AP{sud-africaine} SBC{BG{qui} FV{ne
sont pas produits}} PP{au NP{pays}} , notamment NP{les figues}
séchées , NP{des fruits} comme ENM{NP{les cerises} AP{noires}
et NP{les baies}} , et NP{les haricots} AP{secs} .}
```

6.6.2 Dépendances

Les relations de dépendance concernent des fonctions syntaxiques de base et des relations de nature sémantique dans certains cas d'énumération. Elles sont modélisées sous forme de règles d'extraction ou de déduction.

Extraction

Les règles concernant les fonctions syntaxiques permettent d'extraire des `NP` ou des `PP` arguments du verbe (ou des `AP` attributs d'un verbe copulatif). Les fonctions syntaxiques

extraites sont les suivantes : objet (COMP-OBJ), attribut (VN ou VADJ), argument sous-catégorisé via une préposition (VARG ou VPP).

Pour chaque type de cas, trois règles sont nécessaires : la première extrait le premier NP ou PP de la structure ENM, la deuxième tout NP ou PP après une virgule et la troisième le NP ou PP après une conjonction de coordination si elle existe. Ces arguments verbaux sont extraits si le verbe en question n'a pas d'autres éléments et qui ne font pas partie de l'énumération.

(39) *"Nous ne traiterons que [les compresseurs multi-étages à pistons alternatifs lubriffés, des systèmes de refroidissement par eau, des moteurs d'entraînement diesel, etc.] "*

```
COMP_OBJ_89(traiterons,compresseurs)
COMP_OBJ_90(traiterons,moteurs)
COMP_OBJ_90(traiterons,systèmes)

23>MAX{NP{Nous} FV{ne traiterons} donc BG{que} ENM{NP{les compresseurs}
NP{multi - étages} PP{à NP{pistons}} NP{AP{alternatifs} lubriffés} ,
NP{des systèmes} PP{de NP{refroidissement}} PP{par NP{eau}} , NP{des
moteurs} PP{d' NP{entraînement}} NP{diesel} , etc} .}
```

Dans l'exemple suivant, *"infection"* et *"pneumopathie"* ne sont pas extraits comme VN car le verbe principal a déjà un VADJ.

(40) *" Le syndrome obstructif est responsable de complications respiratoires postopératoires [(infection bronchique, pneumopathie)]."*

```
VADJ_65(est,responsable)

2>MAX{NP{Le syndrome} AP{obstructif} FV{est} AP{responsable}
PP{de NP{complications}} AP{respiratoires} AP{postopératoires}
ENM{ PR{( NP{infection} AP{bronchique} , NP{pneumopathie} )}} .}
```

Déduction

Comme pour les listes, nous avons identifié différents cas de figure pour la déduction de la relation de dépendance RELSEM (entre un nom de la phrase et les différents éléments de la liste qui en dépendent), entre autres :

(A) La relation de dépendance entre un NP sujet du verbe et les items d'une énumération qui le suit :

" Des épreuves fonctionnelles respiratoires sont systématiques : [le V.E.M.S. , la capacité vitale et le rapport V.E.M.S./CV.] "

```
RELSEM(épreuves,V.E.M.S.)
RELSEM(épreuves,capacité)
RELSEM(épreuves,rapport)
```

(B) La relation de dépendance entre un NP argument du verbe et les items d'une énumération qui le suit :

" La nouvelle politique met de l' avant des objectifs principaux , tels que : [l' alimentation , le logement et la salubrité des aliments .] "

```
RELSEM(objectifs,alimentation)
RELSEM(objectifs,logement)
RELSEM(objectifs,salubrité)
```

(C) La relation de dépendance entre un NP (dans un NP avec préposition *"par")* après un verbe passif, et les items d'une énumération qui le suit :

" L' industrie alimentaire sud-africaine est dominée par trois principaux groupes : [CGS Foods , Premier , et Foodcorp.] "

```
RELSEM(groupes,CGS Foods)
RELSEM(groupes,Premier)
RELSEM(groupes,Foodcorp)
```

(D) Comme pour les listes, dans le cas où la tête de la relation sémantique est un NP précédé d'un PP avec la préposition *"de"* on déduit une RELSEM entre les deux noms et les items de l'énumération qui suit :

"Le mandataire assure également l' ensemble des démarches administratives nées à l' importation du véhicule en France : [immatriculation provisoire dans le pays d' origine , passage aux mines , transport .]"

```
RELSEM_113(ensemble,immatriculation)
RELSEM_114(ensemble,transport)
RELSEM_114(ensemble,passage)

RELSEM_128(démarches,immatriculation)
RELSEM_129(démarches,transport)
RELSEM_129(démarches,passage)
```

(E) Dans les cas de phrases avec verbe copulatif et attribut, on extrait une relation sémantique entre le sujet et les items de l'énumération et une autre relation sémantique entre le NP attribut (VN) et les items :

"Les principaux partenaires commerciaux de l'Afrique du Sud sont, outre les pays voisins, [le Royaume-Uni, l'Allemagne, la France et les États-Unis.]"

```
RELSEM(partenaires,Royaume-Uni)
RELSEM(partenaires,Allemagne)
RELSEM(partenaires,États-Unis)

RELSEM(pays,Royaume-Uni)
RELSEM(pays,Allemagne)
RELSEM(pays,États-Unis)
```

Résultat final

Comme pour d'autres phrases de type N2, le résultat final est constitué de la liste de relations de dépendance spécifiques, les relations de dépendance de base et la phrase découpée en différents segments :

(41) *"Les pouvoirs que la loi accorde pour contrer les diverses formes de pollution ou les prévenir sont considérables : [autorisations, permis, avis, rapports, plans, inspections, amendes, ordonnances de fermeture, interventions d'urgence, etc.] "*

```
RELSEM(pouvoirs,autorisations)
RELSEM(pouvoirs,permis)
RELSEM(pouvoirs,inspections)
RELSEM(pouvoirs,plans)
RELSEM(pouvoirs,rapports)
RELSEM(pouvoirs,avis)
RELSEM(pouvoirs,amendes)
RELSEM(pouvoirs,ordonnances)
RELSEM(pourvoirs,interventions)

SUBJ(sont,pouvoirs)
SUBJ(accorde,loi)
COMP_DIR(contrer,formes)
VPP(contrer,de,pollution)
NPP(formes,de,pollution)
NPP(ordonnances,de,fermeture)
NPP(interventions,d',urgence)
VADJ(sont,considérables)

78>MAX{NP{Les pouvoirs} SBC{BG{que} NP{la loi} FV{accorde}} IV{pour contrer}
NP{les AP{diverses} formes} PP{de NP{pollution}} FV{sont} AP{considérables} :
ENM{NP{autorisations} , NP{permis} , NP{avis} , NP{rapports} , NP{plans} ,
NP{inspections} , NP{amendes} , etc} .}
```

6.6.3 Évaluation du module (énumérations)

Nous avons utilisé deux corpus récupérés du Web en août et en octobre 01. Le premier (35.169 mots) appartient aux domaines juridique, scientifico-technique et économique ; le deuxième (19.934 mots) aux domaines juridique, religieux, géographique et psychologique.

Pour les résultats reportés dans la table suivante, la *Précision* correspond au pourcentage d'énumérations correctes (uniquement celles de limites droite et gauche bien repérées) :

	Total énumérations	*Précision*	*Rappel*
Corpus 1	47	74,4 %	89,4 %
Corpus 2	30	83,3 %	93,3 %
Total	77	78,0 %	90,9 %

On observe que, globalement, le taux de précision d'identification des énumérations est inférieur à celui des titres et des listes (96,8 % et 84,5 %, respectivement, *cf.* 5.2.4). Cela confirme la difficulté de repérage de ce phénomène étant donné, d'une part, la variété de configurations et, d'autre part, l'absence d'indices formels homogènes (ponctuation, organisateurs, etc.).

Nous montrons par la suite quelques exemples d'erreur. Dans ce premier cas, l'erreur de marquage de la limite gauche est due à une erreur d'étiquetage (*"s'agissant de"* étiqueté comme une préposition et non pas comme un participe présent)[17] :

(42) *"D'autres décisions sont citées , rendant obsolètes* ⊢ *certains textes réglementaires et législatifs français , s'agissant [des garanties des droits de la défense , du respect de la vie privée , etc .] "*

```
1510>MAX{NP{D'autres décisions} FV{sont citées} , GV{rendant}
AP{obsolètes} ENM{NP{AP{certains} textes} NP{AP{réglementaires
et législatifs} français} , PP{s'agissant des NP{garanties}}
PP{des NP{droits}} PP{de NP{la défense}} , PP{du NP{respect}}
PP{de NP{la vie}} AP{privée} , etc} .}
```

Dans ce deuxième exemple, l'erreur de marquage de la limite droite est due au fait que, pour éviter trop d'erreurs, nos règles de marquage contraignent les énumérations à n'avoir qu'une seule coordination (à moins qu'elle ne soit incluse dans un *chunk* de base comme l'`AP` de l'exemple précédent *"réglementaires et législatifs"*) :

(43) *"Les additifs et les ingrédients alimentaires englobent les grands secteurs suivants : [produits de la minoterie , graines et fruits oléagineux* ⊣ *, gommes et résines , matières à tresser d' origine végétale et graisses et huiles animales ou végétales.] "*

[17] Toujours entre crochets l'énumération existante, un ⊢ ou ⊣ représentant la fausse marque.

```
124>MAX{NP{Les additifs} et NP{les ingrédients} AP{alimentaires}
FV{englobent} NP{les AP{grands} secteurs} suivants : ENM{NP{produits}
PP{de NP{la minoterie}} , NP{graines} et NP{fruits} AP{oléagineux}} ,
NP{gommes} et NP{résines} , NP{matières} IV{à tresser} PP{d' NP{origine}}
 AP{végétale} et NP{graisses} et NP{huiles} AP{animales ou végétales} .}
```

Nous avons remarqué plus d'erreurs pour le marquage de la limite gauche (douze cas) que pour la limite droite (un cas) sur l'ensemble des deux corpus.

La table suivante montre les résultats de l'extraction de quelques dépendances spécialisées pour les énumérations, ainsi qu'une dépendance de base (sujet). `X-ENUM` correspond à des objets, attributs et autres arguments verbaux dans les cas où les arguments sont les items d'une énumération.

	Précision	*Rappel*	*F1*
X-ENUM	100 %	66,6 %	80,0%
RELSEM	61,2 %	74,4 %	67,2 %
SUBJ	94,0 %	92,9 %	93,5 %

Il y a peu de cas de fonctions syntaxiques entre un verbe fini et les items d'une énumération. Cela se justifierait parce que les amorces des listes (séquences introductrices) sont complètes, les items des listes agissant plutôt comme exemples (à la différence des listes).

Pour ces fonctions syntaxiques, les règles sont « prudentes » dans le sens où elles privilégient la précision par rapport au rappel (seule une partie des relations est extraite, mais les relations extraites sont correctes).

En revanche, pour la relation `RELSEM`, on préfère extraire un maximum de cas pour augmenter les probabilités de trouver le bon terme (on tient compte ainsi de la variété de constructions syntaxiques possibles).

Enfin, le taux de précision de la relation `SUBJ` reste élevé : la présence (et le traitement) d'énumérations n'altère pas les résultats pour l'extraction de cette fonction syntaxique.

6.7 Récapitulatif

L'ensemble des grammaires que nous avons réalisées « couvre » environ deux tiers des phrases des corpus de domaines variés. La notion de « couverture » est liée à la fiabilité des analyses : pour ces deux tiers des corpus, notre modèle d'analyseur produit une analyse qui se maintien stable et homogène pour la relation de base sujet et atteint des taux de précision supérieurs à 70 % lors du traitement de certains phénomènes complexes.

Le tableau suivant montre la distribution de types de phrases couvertes par l'analyseur pour un corpus d'exemple (avec 3.813 phrases et environ 108.000 mots) de domaines variés (journaux généraux et spécialisés –*Les Echos*–, corpus scientifiques, corpus juridiques) :

	Nbre de phrases	*Pourcentages*
Total type N1	765	**20 %**
Total type N2	3048	80 %
Total	3813	100 %
Guillemets et parenthèses	642	16,9 %
Listes et énumérations	311	8,1 %
Titres	777	20,4%
Total N2 *traitées*	1730	**45,4 %**
Reste à traiter spécifiquement	1318	34,6 %

Au vu des résultats obtenus lors de l'évaluation des différents modules, la création de grammaires spécialisées améliore le traitement des phrases « complexes ». En effet, des taux de précision supérieurs à 90 % sont maintenus pour le marquage des constituants de base et pour l'extraction de la dépendance de base (sujet). En même temps, une information syntaxique plus précise est extraite grâce à la formalisation de phénomènes jusqu'à présent ignorés par la plupart des analyseurs syntaxiques existants.

Plus précisément, les constituants de base de la grammaire noyau sont analysés avec un taux de précision de 96,1 %. Les constituants spécialisés (`GM`, `AMR`, etc.) sont analysés avec des taux situés entre 70 % et 87 %. Pour l'extraction de dépendances, le tableau suivant montre le taux de précision de la dépendance de base sujet ainsi que les moyennes (pondérées) des différentes relations de dépendance spécialisées (`NPR`, `RELSEM`, etc.).

	Dépendance de base	*Dépendances spécialisées*
Gram. Noyau	99,1 %	-
Gram. Ponctuation	94,8 %	70,3 %
Gram. Listes	90,4 %	86 %
Gram. Énumérations	94 %	80,6 %
Gram. Titres	100 %	93 %

Nous pouvons conclure que, pour les phrases N1 et pour les phrases N2 présentant des phénomènes modélisés par nos grammaires, notre modèle d'analyseur produit une analyse linguistique fine. En effet, des évaluations de la dépendance `sujet` montrent qu'il n'y a pas de variations en qualité selon les différents phénomènes. On obtient donc une analyse globalement fiable dont la qualité reste stable à travers l'hétérogénéité des corpus.

6.8 Résumé

Dans ce chapitre, nous avons présenté les différentes grammaires spécialisées de notre modèle d'analyseur syntaxique. Comme nous l'avons vu, il s'agit de quatre modules grammaticaux (grammaire pour le traitement de la ponctuation, des titres, des listes et des énumérations) qui s'appliquent sur des phrases contenant des phénomènes nécessitant un traitement syntaxique précis.

Les différentes évaluations menées montrent que les taux de précision et de rappel sur les structures et dépendances créées présentent des taux supérieurs à 90 % dans la plupart des cas sans pour autant faire baisser la précision de l'analyse de structures et dépendances de base.

À la lumière des résultats obtenus, la création de plusieurs grammaires modulaires et reconfigurables s'avère une solution possible pour l'amélioration du traitement linguistique précis et homogène de textes tout venant.

Troisième partie

Lexicalisation des grammaires de dépendances

Chapitre 7

Apprentissage de patrons de cooccurrence

7.1 Introduction

Après avoir présenté notre travail en ce qui concerne la création de pré-traitements et de grammaires, nous allons dans la suite décrire une méthode qui permet d'améliorer les sorties de l'analyseur pour le traitement des ambiguïtés structurelles, spécialement le rattachement prépositionnel. Cette méthode combine les grammaires symboliques existantes avec des techniques d'apprentissage et utilise le Web comme ressource principale pour l'extraction d'informations lexicales. La combinaison d'une description structurale riche avec des informations statistiques rend notre modèle d'analyseur hybride, à l'instar des approches de [Sri97] et [BM97].

De façon générale, ce chapitre présente la méthodologie permettant l'acquisition automatique de ces informations lexicales (des *patrons de cooccurrence syntaxique*) et le calcul de leurs poids, alors que le chapitre suivant décrit l'utilisation de ces informations pour améliorer les sorties de l'analyseur (levée d'ambiguïtés structurelles concernant le rattachement prépositionnel).

Nous nous sommes proposé d'améliorer les résultats du rattachement prépositionnel dans le cadre de l'analyseur syntaxique présenté dans les chapitres précédents. Pour ce faire, nous avons développé une méthode d'apprentissage non supervisé à partir d'un corpus annoté par l'analyseur.

Avant de montrer l'ensemble de cette méthode, nous faisons d'abord en section 7.2 un survol de différentes approches existantes pour le traitement du rattachement prépositionnel, spécialement des méthodes statistiques.

Dans la section 7.3, nous exposons les grandes lignes de notre approche, que l'on peut décomposer en deux grandes étapes :

- extraction de dépendances liées au rattachement prépositionnel avec une grammaire à base de règles et stockage sous forme de patrons de cooccurrence pondérées ;
- utilisation de ces patrons pour la levée d'ambiguïtés de rattachement.

Les sections suivantes détaillent la première étape (la deuxième fait l'objet du prochain

chapitre). Ainsi, la section 7.4 décrit nos grammaires pour l'extraction de dépendances et la section 7.5 la construction de la base de patrons à partir d'un très grand corpus, avec le calcul des poids de rattachement (mesure d'estimation de la probabilité de rattachement). Un résumé en section 7.6 synthétise les apports de cette première étape de la méthode que nous proposons.

7.2 Approches existantes

En général, on distingue les approches linguistiques des approches statistiques. Les différentes méthodes étudient principalement le cas du rattachement prépositionnel en anglais pour la structure `VP NP PP` (nous avons observé peu de travaux concernant d'autres cas ambigus comme `NP NP PP` [BR94], `NP AP PP` ou encore `VP NP PP PP` [MCB97]). Pour ce qui est d'autres cas d'ambiguïté de rattachement, [Jac97] propose une méthode de désambiguïsation pour des structures `(NP3 NP2) NP1` et `NP3 (NP2 NP1)` en anglais, basée sur des informations statistiques collectées dans des grands corpus.

7.2.1 Approches linguistiques

Il existe des approches linguistiques qui tentent de résoudre le problème du rattachement prépositionnel en utilisant *uniquement* des propriétés structurelles sans apport d'information lexicale. Les principes plus courants sont :

- **l'attachement minimal** (*minimal attachment*, [Fra78]) : un constituant tend à être attaché de façon à impliquer le nombre minimum de nœuds non-terminaux.

- **l'association à droite** (*late closure* ou *principle of right association*, [Kim73]) : un constituant tend à être attaché au constituant immédiatement à sa droite.

En cas de conflit lors du rattachement prépositionnel, l'attachement minimal prédomine. Les deux principes font des prédictions contraires car le premier prédit le rattachement au verbe et le deuxième le rattachement au nom (toujours dans une structure `VP NP PP`). Les résultats obtenus ne sont guère satisfaisants, d'autant plus que la couverture de phénomènes est très réduite (55 % des cas, [HR93]).

D'autres méthodes utilisent des informations linguistiques plus riches pour résoudre le problème du rattachement prépositionnel. Par exemple, [Sch95] propose la préférence de rattachement des arguments au rattachement des modifieurs, [JB87] l'utilisation de restrictions de sélection basées sur des classes sémantiques. Cependant, ces méthodes se heurtent à un problème non négligeable : l'absence de ressources complètes et facilement accessibles (réseaux sémantiques, dictionnaires de rection, etc.) [Vol01a].

7.2.2 Approches statistiques

Dans les méthodes statistiques on distingue les approches supervisées et les non supervisées selon que le corpus d'entraînement est manuellement contrôlé -annoté- (méthodes supervisées) ou bien annoté automatiquement ou pas annoté du tout (méthodes non supervisées).

Méthodes supervisées

La littérature fait état de plusieurs modèles :

- **apprentissage par transformations** (*Transformation-based learning*, [BR94]) : l'algorithme crée des règles à partir d'exemples du corpus d'entraînement à partir d'une règle initiale "simple" ; lors de l'apprentissage, des règles de correction s'appliquent de façon ordonnée.

- **modèle de l'entropie maximale** (*Maximal entropy model*, [RRR94], [RRR97]) : il cherche à calculer la probabilité de distribution de la décision de rattachement en utilisant uniquement l'information contenue dans le `VP` et avec des valeurs obtenues par l'application de fonctions définies initialement sur le corpus d'entraînement.

- **apprentissage basé sur la mémoire** (*Memory-based learning*, [ZDV97]) : cette méthode stocke des exemples en mémoire et les généralise en utilisant des mesures de similarité *intelligentes* ; elle est fondée sur les régularités des cooccurrences.

- **modèle « backed-off »** (*Backed-off model*, [CB95]) : ce modèle détermine la probabilité de rattachement à partir des fréquences de rattachement obtenues dans le corpus d'entraînement.

- **méthode « boosting »** (*Boosting*, [ASS96]) : cette technique combine des règles simples de façon ordonnée pour produire des règles précises de classification.

La table 1 récapitule les différents taux de précision obtenus par ces approches (uniquement pour la structure `VP NP PP`) sur un même jeu de données en anglais : des phrases extraites au hasard du Wall Street Journal[1].

Méthode	Auteurs	Précision
transformation-based learning	[BR94]	80,0 %
maximal entropy model	[RRR94]	81,6 %
memory-based learning	[ZDV97]	84,4 %
backed-off model	[CB95]	84,5 %
boosting	[ASS96]	84,5 %
annotateurs humains (1)		88,2 %
annotateurs humains (2)		93,2 %

Table 1. Résultats des différentes méthodes d'apprentissage supervisé.

Certains approches combinent aprentissage supervisé et informations linguistiques. [WF96] utilisent des règles qui encodent des informations lexicales ou sémantiques (par exemple, un PP indiquant « temps » ou « lieu » est plutôt rattaché au verbe) et obtiennent une moyenne de 86,9 % de taux de précision de rattachement ; [SN97] utilisent le réseau WordNet et obtiennent un taux de précision de 88 %. Leurs approches démontrent

[1] Les résultats obtenus par les humains sont rapportés par [CB95], (1) en utilisant uniquement les informations de la structure `VP NP PP`, (2) en utilisant le contexte de toute la phrase.

que l'utilisation d'information linguistique améliore la résolution du rattachement prépositionnel.

Pour ce qui est de l'attachement multiple, il existe peu de travaux décrits dans la littérature. Dans [MCB97], les auteurs proposent une méthode *backed-off* généralisée (pour l'anglais) pour des configurations telles que `VP NP PP1 PP2` et `VP NP PP1 PP2 PP3`. Leur méthode s'avère positive dans le cas d'un seul rattachement `VP NP PP1` (84,3 %, résultat similaire à celui de [CB95]) mais ces résultats diminuent considérablement pour le rattachement multiple (69,6 % pour `PP2` et 43,6 % pour `PP3`).

Les auteurs justifient cette baisse par la forte augmentation du nombre de configurations à prendre en compte ainsi que par le problème de la quantité de données (les structures de `PP` multiples sont moins fréquentes dans les corpus et contiennent plus de mots, ce qui demande un espace de solutions beaucoup plus grand).

Méthodes non supervisées

Les méthodes non supervisées exploitent les régularités des corpus tout venant ou annotés automatiquement [Vol01a]. Le bruit lors de la récupération des données (informations non pertinentes) est compensé par la quantité de matériel collecté. Ces approches calculent les fréquences d'apparition d'une préposition à côté d'un verbe ou nom donné (collocations).

L'approche plus classique est celle des **associations lexicales** (*lexical association scores*, [HR93]). Cette méthode estime la probabilité des associations des prépositions par rapport à un nom ou à un verbe (d'après leurs distributions dans les corpus annotés par un parseur). Le taux de précision rapporté est de 80 %.

Les taux de précision d'autres techniques se situent entre 81,9 % et 84,3 % [Vol01a].

7.2.3 Approches hybrides

D'autres approches généralisent en un modèle statistique plusieurs des techniques décrites dans les méthodes supervisées tout en intégrant des sources d'information différentes (informations lexicales comme les rections ou des ressources sémantiques). [GC01] obtiennent des résultats pour le français (corpus du journal *Le Monde*) qui se situent autour de 85 % de précision pour la structure `VP NP PP` et autour de 74 % en incluant aussi d'autres structures ambiguës comme le rattachement au nom dans une configuration telle que `NP NP PP`.

Pour l'anglais, la combinaison d'une méthode statistique avec l'utilisation d'informations linguistiques (par exemple, les classes sémantiques dans WordNet) permet d'obtenir des meilleurs résultats par rapport aux méthodes statistiques décrites plus haut (récapitulatif Table 1). En effet, pour la résolution du rattachement prépositionnel dans la configuration `VP NP PP`, la méthode de [SN97] obtient un taux de précision de 88 %.

Ce bref survol sur les méthodes existantes nous amène à conclure sur les deux idées suivantes :

(a) dans la plupart des cas, les méthodes utilisées pour résoudre le rattachement prépositionnel sont des approches supervisées, car il s'avère plus efficace de comparer les résultats avec des corpus où il y a eu de l'intervention humaine que des corpus traités uniquement de façon automatique ;

(b) quelle que soit la technique, le taux de précision pour la désambiguïsation du rattachement atteint au plus 84,5 % en se limitant à la structure la plus étudiée (`VP NP PP`), et elle descend de façon significative dès que des ambiguïtés de rattachement de types différents sont prises en compte.

7.3 Aperçu de notre méthode

Nous présentons dans cette section les lignes générales de notre méthode. Avant cela, nous faisons tout d'abord quelques précisions en ce qui concerne la terminologie utilisée. Par la suite, nous montrons à quel niveau de l'analyse se situe notre travail (au sein de l'architecture de l'analyseur) et nous présentons les corpus initiaux et les critères d'évaluation utilisés.

7.3.1 Terminologie

Par souci de clarté, nous définissons dans cette section plusieurs notions importantes pour notre approche.

Premièrement, la grammaire de dépendances extrait des relations de dépendance entre un `PP` et un autre syntagme dont il dépend (`VP`, `NP` ou `AP`). On définit une relation de dépendance concernant le rattachement prépositionnel comme suit :

> « Dépendance » : structure sous la forme A(X,Prép,N) où A correspond au nom de la dépendance (A pour *attachment*), X est un nom, un verbe ou un adjectif, Prép une préposition et N un nom.

La tête syntaxique de la relation est toujours `X` et l'élément dépendant `N`, par exemple `A(durée,de,vie)` dans *Sa durée de vie est de 10 ans*. Les dépendances extraites sont transformées en patrons de cooccurrence et stockées dans un base.

La notion de patron de cooccurrence est, dans ce contexte, plus large que la notion classique de rection ou valence. En effet, la linguistique « classique » définit une *rection* comme un élément (généralement un verbe, mais aussi un nom ou un adjectif) suivi de la préposition qui introduit l'argument régi. Un *argument* est toujours un complément obligatoire ou essentiel [Gre93].

Dans notre approche, nous ne nous intéressons pas à la fonction syntaxique de l'élément rattaché (nous ne faisons pas la différence entre « argument » ou « circonstant », c'est-à-dire que nous ne tenons pas compte du caractère obligatoire ou optionnel de l'élément rattaché. Même si cette différence est bien établie dans la littérature linguistique et généralement admise, l'identification automatique est parfois difficile [BF00], [FF02]. Dans ce travail, nous sommes donc uniquement intéressée par la résolution du rattachement ambigu (à quelle tête se rattache un syntagme prépositionnel), qu'il soit facultatif

ou obligatoire. Nous généralisons alors « argument » et « circonstant » avec la notion de patron de coocurrence :

> « Patron de cooccurrence syntaxique » : structure formée des éléments X (nom, verbe ou adjectif) et Prép (préposition) provenant d'une relation de dépendance de type A(X,Prép,N).

Finalement, les règles de la grammaire permettant l'extraction de dépendances (liées au rattachement prépositionnel) peuvent être regroupées en deux types, selon leurs performances (évaluées automatiquement, voir section 7.4). Nous distinguons ainsi les règles « fiables » des règles « non fiables » :

> « Règle fiable » : règle de grammaire qui extrait des relations de dépendance A(X,Prép,N) avec un taux de précision d'au moins 93 % (*cf.* section 7.4.2).

7.3.2 Vue générale

Comme nous l'avons indiqué plus haut, nous nous sommes proposé d'améliorer les résultats du rattachement prépositionnel, quelle que soit sa configuration, dans le cadre de l'analyseur syntaxique présenté dans les chapitres précédents.

Or, la résolution d'ambiguïtés de rattachement demande d'aller au-delà des informations purement syntaxiques. De ce fait, notre approche comprend, globalement :

- l'ajout d'informations lexicales (patrons de cooccurrence) aux grammaires de dépendances ;
- la combinaison d'informations statistiques (apprentissage non supervisé, mesures de fréquence, etc.) avec des grammaires.

Plus concrètement, nous construisons une base de patrons de cooccurrences à partir d'une première analyse syntaxique d'un très grand corpus. Cette analyse est faite par une grammaire construite manuellement. Par la suite, des mesures de fréquence sont ajoutées à chaque patron et calculées à partir de ce corpus. Ces mesures (poids) sont fondées sur la *cooccurrence syntaxique* et non pas sur une *cooccurrence* purement *textuelle*. Ainsi, nous faisos l'hypothèse que des mots cooccurrents dans une même relation syntaxique produite par l'analyseur lors d'une première analyse sont plus « fiables » que des cooccurrences dans des corpus tout venant, car l'analyseur élimine des cas où un lien syntaxique serait impossible (par exemple, entre deux propositions différentes).

Ces informations lexicales et statistiques sont alors utilisées pour lever les ambiguïtés de rattachement lors des analyses suivantes. La grammaire de notre modèle d'analyseur est ainsi lexicalisée et comprend une étape supplémentaire d'analyse dédiée à la désambiguïsation structurelle. La figure 7.1 schématise l'architecture générale du modèle.

La méthode proposée se divise en deux grandes étapes. La première étape consiste à construire d'une base d'informations lexicales (patrons de cooccurrence pondérés) et est suivie d'une étape de levée d'ambiguïtés.

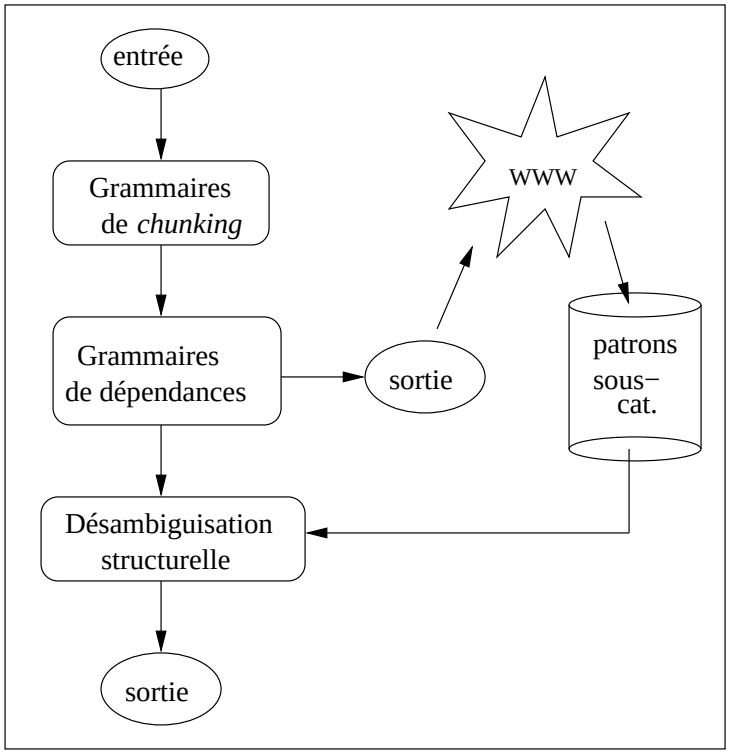

FIG. 7.1 – Architecture générale de notre approche.

Initialement, notre analyseur extrait des dépendances d'un texte en entrée (le calcul de ces dépendances est fondé sur le résultat de l'analyse des grammaires de chunking). Ces dépendances contiennent les têtes des syntagmes avec les lemmes et les numéros de position dans la phrase, par exemple :

```
A(<égal:20>,<à:21>,<angle:23>)
A(<juger:7>,<de:8>,<efficacité:10>)
A(<avoir:8>,<sur:12>,<compte:14>)
A(<impact:10>,<sur:12>,<compte:14>)
A(<résultat:23>,<de:24>,<entreprise:26>)
A(...)
```

Nous transformons chaque dépendance en un patron de cooccurrence (seuls les éléments `X` et `Prép` sont conservés) et nous stockons ces patrons dans une base avec une mesure de la probabilité de rattachement calculée pour chaque patron (nous donnons les détails pour le calcul de ce score dans la section 7.5.5). La base se présente sous la forme suivante (la première colonne correspond à cette mesure de probabilité de rattachement, la deuxième au patron de cooccurrence) :

```
0.6786 égal|à
```

```
0.1096 juger|de
0.0118 avoir|sur
0.0626 impact|sur
0.2976 résultat|de
0..... ...
```

Lors de la levée des ambiguïtés, nous avons créé un algorithme qui utilise ces informations pour choisir le rattachement le plus probable parmi deux ou plusieurs dépendances partageant un même élément dépendant (`N`) et extraites lors d'une première analyse. Ainsi, pour la phrase en exemple, notre analyseur extrait dans un premier temps les dépendances suivantes :

(1) *"La méthode d'amortissement aura un impact significatif sur les comptes"*

```
A_MF1(<méthode:4>,<de:5>,<amortissement:6>)
A_MF2(<avoir:8>,<sur:12>,<compte:14>)
A_MF2(<impact:10>,<sur:12>,<compte:14>)
A_MF2(<significatif:11>,<sur:12>,<compte:14>)
```

La première dépendance est produite par une règle fiable : elle est donc validée directement (elle est gardée comme correcte et reproduite lors de la sortie finale). Parmi les trois rattachements en conflit, seulement `A(impact,sur,compte)` est validé car son poids est supérieur à celui de `A(avoir,sur,compte)` ($0.0626 > 0.0118$), et celui de `A(significatif,sur,compte)` est inexistant dans la table donc équivalent à zéro.

Cette méthode améliore les résultats de la grammaire d'extraction de dépendances (comme nous l'expliquons en détail à la section 8.2.3) grâce à l'information lexicale acquise à partir du corpus et pondérée par la suite avec des mesures d'estimation de la probabilité de rattachement.

La décision d'acquérir l'information lexicale (les patrons de cooccurrence) à partir d'un corpus et non pas à partir d'un dictionnaire (type Dubois [DDC97] ou AlethDic [Gsi93]) est motivée par deux raisons.

D'une part, les patrons présents dans ces dictionnaires concernent généralement les « arguments » (éléments régis) or on retrouve souvent un manque de rigueur pour distinguer clairement « argument » et « circonstant » dans le contexte de ces dictionnaires. De plus, il n'existe pas à l'heure actuelle pour le français une base lexicale complète et facilement accessible contenant des informations sur verbes, noms et adjectifs [BF00].

D'autre part, les dictionnaires n'incluent pas des mesures de poids pour chaque rection. Il est donc impossible de savoir si une rection est fréquente ou pas, ou bien entre deux rections avec une même tête syntaxique, laquelle des deux est la plus fréquente. Cette information est cruciale pour notre approche. (Dans des cas d'ambiguïté de rattachement, quelle rection choisir s'il n'y a pas d'information statistique ?).

Ces deux raisons nous ont amenée à constituer notre propre base de patrons de cooccurrence pondérés, selon la méthode que nous décrivons ci-dessous.

7.3.3 Annotation des corpus initiaux

Pour la mise en œuvre de notre travail, nous avons utilisé deux corpus différents (provenant du Web) des domaines économique, juridique et scientifique (Annexe E).

Le premier corpus, que nous appelons « corpus A » a été utilisé pour le développement des différentes grammaires visant à obtenir un ensemble de règles le plus approprié pour nos objectifs (*cf.* 7.4). Ce corpus est constitué de 89 phrases[2] et de 2.493 mots.

Le deuxième corpus, appelé « corpus B », est le corpus de référence utilisé pour valider la grammaire *de base* (*cf.* 7.4.2) ainsi que les différentes expériences d'apprentissage (*cf.* 8.2 et 8.3). Il est constitué de 337 phrases et de 9.339 mots.

Ces deux corpus ont été semi-automatiquement annotés : ils ont été analysés par nos grammaires noyau et spécialisées pour le *chunking* et par une grammaire primitive d'extraction de dépendances. Nous avons corrigé manuellement ces analyses[3] dans le but d'obtenir deux corpus de référence pour les tâches que nous nous sommes données (le corpus A pour la validation de grammaires, le corpus B pour la validation de la grammaire *de base* et de la méthode d'apprentissage).

L'ensemble de ces deux corpus constitue un corpus de 11.628 mots et 426 phrases avec 1.381 dépendances validées.

7.3.4 Critères d'évaluation

Pour l'évaluation des résultats obtenus lors de la création de grammaires de dépendance et lors de l'application de l'algorithme d'apprentissage pour la levée des ambiguïtés nous avons utilisé les deux mesures classiques suivantes (*cf.* 2.3) : la précision (`P`) et le rappel (`R`). Nous avons aussi utilisé une moyenne pondérée de ces deux mesures, la moyenne `F1`, qui atteint un maximum de 100 % quand les taux de précision et de rappel sont aussi de 100 %.

Ces trois mesures classiques sont utiles pour l'évaluation des performances des différentes grammaires ainsi que pour la comparaison des résultats obtenus uniquement avec la grammaires symbolique et à la suite de l'apprentissage.

7.4 Grammaires pour l'extraction de dépendances

Dans cette section nous présentons l'ensemble de grammaires que nous avons créées. La définition de ces grammaires constitue une étape préalable à l'introduction d'information lexicale. En effet, ce sont les sorties de l'analyseur produites par ces grammaires qui seront utilisées par la suite pour l'acquisition automatique des informations lexicales.

7.4.1 Grammaires initiales

L'idée étant de construire une base de patrons de cooccurrence à partir des dépendances extraites lors d'une première analyse, nous avons défini tout d'abord deux grammaires avec des heuristiques différentes. L'extraction de dépendances liées au rattache-

[2] Pour rappel, la notion de *phrase* est ici élargie : elle inclut listes et titres.

[3] Le volume de travail humain représente, approximativement, une semaine à temps complet.

ment prépositionnel est non déterministe car notre analyseur calcule tous les rattachements possibles selon les heuristiques définies dans les règles de la grammaire.

Les deux grammaires initiales, préalables à l'introduction de techniques d'apprentissage, sont :

Gram. G0 : priorité au rappel.

Gram. G1 : priorité à la précision.

La première grammaire favorise le rappel, c'est-à-dire que les règles qui la constituent présentent peu de contraintes dans le but de fournir un maximum de relations de dépendance. En revanche, la deuxième grammaire limite le nombre de relations extraites et favorise ainsi la précision des dépendances produites.

Leurs différences sont significatives :

Grammaire	Précision	Rappel	F1	Déps. extraites
G0	41,88 %	**94,28**	58,00 %	2.438
G1	**87,49 %**	77,04 %	**81,93 %**	964

On peut conclure que la grammaire G1 s'avère une meilleure grammaire symbolique pour l'extraction de dépendances liées au rattachement prépositionnel.

7.4.2 Grammaire de base

En vue de l'apprentissage de patrons de cooccurrence, nous faisons l'hypothèse que, plus la base de patrons sera riche, plus l'algorithme de levée des ambiguïtés aura des données pour résoudre les rattachements. Ainsi, pour obtenir une base importante, nous avons besoin d'un maximum de relations de dépendances extraites d'un corpus.

Or, en prenant la grammaire G1, on se rend compte qu'elle n'est pas très efficace pour valider nos hypothèses car, bien qu'elle ait un taux de précision assez bon, son rappel est bas : le nombre de dépendances extraites d'un corpus n'est donc pas assez significatif.

Par conséquent, et dans le but d'optimiser le résultat de G1 avec l'apprentissage, nous avons considéré nécessaire de créer une nouvelle grammaire, que nous appelons G2, et qui sert « de base » pour valider notre méthode (voir un échantillon des règles de G2 dans l'annexe D.6).

Cette grammaire **G2** intègre :

- des règles de la grammaire G1 (uniquement les règles « fiables ») ;
- les règles de la grammaire G0 qui s'appliquent uniquement s'il n'y a pas de rattachement extrait par les règles précédentes.

Ainsi, parmi les règles incluses dans la grammaire G1, et après évaluation automatique de leurs taux de précision, nous avons identifié un ensemble de règles avec un taux de précision supérieur à 93 %.

Voici le résultat pour chaque règle de G1 pour le corpus B sur un total de 1.084 dépendances[4] :

```
   Précision pour chaque règle :
Règle 1 (235 :24) : P= 89.79 %
Règle 2 (3 :0) : P= 100.00 %
Règle 3 (3 :0) : P= 100.00 %
Règle 4 (3 :0) : P= 100.00 %
Règle 5 (494 :31) : P= 93.72 %
Règle 6 (44 :1) : P= 97.73 %
Règle 7 (1 :0) : P= 100.00 %
Règle 8 (33 :7) : P= 78.79 %
Règle 9 (9 :3) : P= 66.67 %
Règle 10 (84 :49) : P= 41.67 %
Règle 11 (50 :13) : P= 74.00 %
```

D'après ces résultats, nous considérons les règles ayant un taux de précision supérieur à 93 % comme des règles *fiables* et les autres règles comme des règles *moins fiables*.

En se basant sur les résultats obtenus sur le corpus B (mais en prenant en compte aussi les résultats pour le corpus A), on identifie comme fiables les règles R2, R3, R5, R6, R7.

À noter que la règle R4 donne 100 % de précision (3 dépendances extraites) pour le corpus B, mais seulement 50 % (2 dépendances erronées sur 4) dans le corpus A. C'est pourquoi on a décidé de la considérer comme plutôt moins fiable.

Les dépendances produites par des règles fiables ont le trait `MF1` associé. Parmi ces règles, il y a des cas de configurations très sûres, comme entre un début de phrase et un verbe principal (`NP PP VP`) ou entre un verbe principal et la fin de phrase (`VP PP.`).

Un autre cas est celui de la configuration `NP` suivi d'un `PP` avec la préposition *de*, modélisé avec la règle :

```
if (~A(?,#2,#3))
A[mf1=+](#1,#2,#3) =
NP[start:~]{?*,#1[last:+,num:~,time:~,pron:~]};
PP[start:~]{?*,NP{?*,#1[last:+,num:~,time:~,pron:~]}};
BG{?*,pron#1[last:+]},
?*[noun:~,verb:~,prep:~,pron:~,coord:~,form:~fopar,
form:~fcpar,form:~f2pts,conjque:~,punct:~],
PP[fnpp=+]{?*, #2[prep:+,sfde:+], ?*,NP{?*,#3[last:+]}}.
```

Cette règle indique qu'un nom inclus dans un `NP` ou `PP` initial (avant le verbe fini) se rattache à un `PP` ayant la préposition *de* si, entre ces deux constituants, il n'y a pas un autre nom, ou un verbe fini, ou une préposition, ou une coordination, ou une conjonction *que*, ou bien un signe de ponctuation quelconque.

[4] À coté de chaque règle, il y a le nombre d'occurrences extraites et le nombre d'erreurs : `(235 :24)` signifie 235 dépendances extraites avec la règle donnée, dont 24 erreurs.

Dans l'exemple suivant, cette règle extrait deux dépendances :

(2) *"La présence des Verts au Conseil de Paris est donc une chance pour la nouvelle majorité municipale."*

```
A_MF1(<conseil:5>,<de:6>,<Paris:7>)
A_MF1(<présence:1>,<de:2>,<vert:3>)
```

La grammaire G2 contient en tout cinq règles fiables provenant de G1 qui couvrent environ 47 % des dépendances extraites. À ces règles, nous avons ajouté les règles provenant de la grammaire G0, qui donnent priorité au rappel (voir figure 7.2).

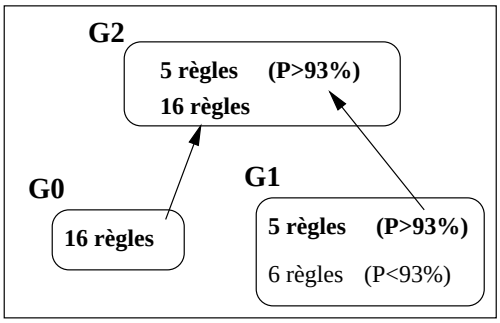

FIG. 7.2 – Composition de la grammaire G2.

L'ensemble des règles provenant de G0 ne s'appliquera pour un PP donné que si ce PP n'a pas déjà été rattaché à une tête syntaxique par le biais d'une règle fiable.

Un exemple de ce type de règle non fiable est celui de l'extraction d'une dépendance entre un verbe (conjugué, participe passé ou présent) et un PP, avec la condition que l'élément à rattacher ne soit pas extrait par une règle fiable et la seule contrainte qu'il n'y ait pas d'autre verbe, conjonction ou deux points entre le verbe et le PP.

```
if (~A[mf1:+](?,?,#3) )
A[mf2=+](#1,#2,#3) =
        FV{?*,#1[last:+]};IV{?*,#1[last:+]};GV{?*,#1[last:+]};
SBC{?*,FV{?*,#1[last:+]}};verb#1[partpas:+],
        ?*[form:~f2pts,verb:~,prep:~,scbegin:~,coord:~],
        PP[fonc:~,fvpp=+,fadjpp:~]{?*[prep:~],#2[prep:+],?*,NP{?*,#3[last:+]}}.
```

Cette règle a une précision que nous avons estimée à 70,31 % sur le corpus B. Elle donne de bons résultats dans des exemples comme le suivant :

(3) *Le Système Odyssey est constitué d' un réseau de 12 satellites d' orbite moyenne, environ 10000 Km, et de 7 stations terrestres .*

```
A_MF2(<constituer:4>,<de:5>,<réseau:7>)
```

En revanche, elle extrait aussi des dépendances incorrectes :

(4) *La partie apparente du contenu du colis ou du lot doit être représentative de l'ensemble .*

```
A_MF2(<être:11>,<de:13>,<ensemble:15>)
```

Nous faisons l'hypothèse que ce type de dépendances incorrectes pourra être corrigé grâce à l'application de notre méthode d'apprentissage.

Les résultats globaux de la grammaire G2 sont les suivants :

Grammaire	Précision	Rappel	F1	Déps. extraites
G2	71,34 %	92,10 %	**80,40 %**	1.413

La moyenne F1 n'est pas très différente de celle de la grammaire G1 mais le taux de rappel est bien supérieur, ce qui est important pour la construction d'une base de patrons de cooccurrence la plus riche possible.

7.5 Construction d'une base de patrons de cooccurrence

Comme nous l'avons écrit plus haut, nous faisons l'hypothèse qu'une base de rections de taille importante constituera un espace de solutions plus grand pour la levée d'ambiguïtés. Il est donc nécessaire d'avoir un corpus initial de grande taille pour augmenter le nombre d'échantillons utilisés pour calculer les probabilités. Nous nous sommes ainsi proposé d'utiliser le Web.

7.5.1 Utilisation d'un très grand corpus : le WWW

L'utilisation du World Wide Web comme grande base d'exemples pour différentes tâches liées au traitement automatique est une idée exploitée depuis peu : pour la traduction de noms composés [Gre99], pour l'acquisition d'entités nommées [JB00], pour la désambiguïsation de relations liées au rattachement prépositionnel [Vol01b][5] et [Leb02].

Étant donné la taille de cette ressource, nous allons l'utiliser pour en extraire des patrons de cooccurrence en faisant l'hypothèse que la taille des échantillons sera suffisamment grande pour améliorer les résultats globaux de la grammaire.

[5] Le travail de Volk ([Vol00], [Vol01a], [Vol01b]) sur le rattachement prépositionnel en utilisant le Web comme corpus a comme points en commun avec notre approche l'idée générale d'utiliser une base de rections avec un poids associé (mesure de la fréquence de cooccurrence). Toutefois, il existe une différence importante avec notre travail : le corpus n'est pas collecté à partir d'un premier corpus ayant des relations de dépendance et les fréquences sont calculées directement à partir des résultats obtenus par le moteur de recherche (fréquences données par Altavista) et seulement pour la configuration `VP NP PP`. Volk utilise donc des *cooccurrences textuelles* et non pas des *cooccurrences syntaxiques*. Les résultats obtenus pour la configuration mantionnée sont d'environ 75 % en ce qui concerne le taux de précision.

7.5.2 Préparation des requêtes

Pour la récupération de corpus provenant du WWW, nous avons tout d'abord créé un fichier de requêtes à partir de toutes les relations `MF2` à valider extraites lors de l'analyse du corpus B. La figure suivante schématise notre démarche globale :

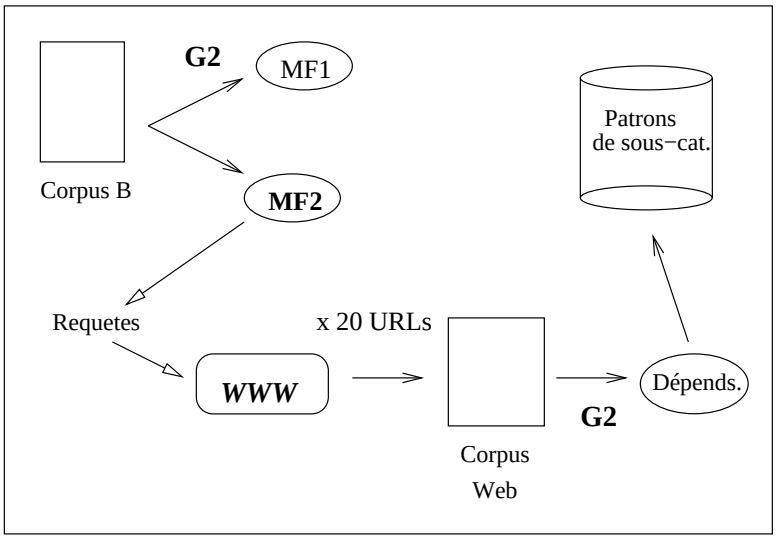

FIG. 7.3 – Apprentissage avec le WWW.

Le corpus initial est analysé avec la grammaire G2. Les dépendances obtenues avec des règles dont la précision est inférieure à 93 % (`MF2`) sont transformées en requêtes et soumises au Web. On s'assure ainsi de trouver, si elles existent, des occurrences de ces relations sur des corpus réels.

On estime que les relations extraites du corpus initial avec des règles fiables (`MF1`) sont déjà correctes ; le but est donc d'améliorer les dépendances produites par des règles moins précises, c'est-à-dire, d'éliminer les dépendances incorrectes en choisissant uniquement les bons rattachements. L'ensemble de ces dépendances moins fiables constitue ainsi le point de départ de notre corpus d'apprentissage.

Nous collectons alors une vingtaine de pages par requête et l'ensemble de ce processus nous permet de créer un corpus de grande taille que nous appelons « corpus web ». Ce corpus est analysé avec la même grammaire G2 dans le but d'extraire un nombre élévé de relations de dépendance qui constitueront notre base d'exemples.

Chaque requête obtenue à partir d'une relation `MF2` contient les trois occurrences exactes de la relation extraite et non pas les lemmes (sans les numéros de position) car aucun des moteurs de recherche existants ne permet des traitements linguistiques sur les termes de la requête (toutes les variantes morphologiques pour un lemme donné).

Nous avons utilisé le moteur de recherche `Altavista` (`www.Altavista.com`) avec l'option de recherche régulière qui permet de formuler des requêtes complexes à l'aide d'opérateurs booléens entre les termes. L'un de ces opérateurs, `NEAR`, restreint la recherche à

des documents où les deux termes apparaissent séparés par un maximum de dix mots. De plus, les parenthèses permettent de regrouper des cooccurrences.

Les dépendances à valider ont été donc transformées selon la syntaxe suivante :

(`X Prép`) `NEAR N`

Quelques exemples sont :

```
(nombre d') NEAR entreprises
(appuyant sur) NEAR réglementation
(accordées aux) NEAR emplois
(juger de) NEAR efficacité
```

Nous avons cherché des URLs qui contiennent `X` et `Prép` contigus suivis du mot `N` avec n mots entre `Prép` et `N` ($n < 10$). Nous sommes consciente que ces requêtes ne donnent pas toujours les résultats attendus, mais cette difficulté est compensée par la grande quantité de données obtenues.

Ainsi, si on cherche à valider la relation `A_MF2(équipements,pour,automobile)`, à partir de la requête `(équipements pour) NEAR automobile` on obtiendra des documents contenant la liste d'exemples suivants (parmi lesquels, seulement les deux premiers seront valides pour nos expériences) :

Document 1 ... **équipements pour automobile**s ...
Document 2 ... **équipements pour** l'**automobile** ...
Document 3 ... **équipements pour** la maison ou les constructeurs **automobile**s ...
Document 4 ... **équipements pour** garages et stations services. Salon international de l'**automobile** ...

Certains des documents obtenus ne sont pas pertinents pour notre tâche (les documents 3 et 4 dans l'exemple), mais nous faisons l'hypothèse que ce problème de pertinence sera pallié par le nombre de bonnes occurrences obtenues dans la grande collection de documents.

7.5.3 Création d'un grand corpus d'apprentissage

À partir des 869 requêtes obtenues après transformation des 869 dépendances `MF2` extraites par la grammaire G2 du corpus B, nous avons collecté sur le Web une vingtaine d'URLs par requête. Pour cela, nous avons utilisé un ensemble de scripts *perl* combinés avec la commande d'Unix `wget`.

Nos scripts permettent de transformer chaque requête en une requête HTML qui sera lancée sur le Web. Les vingt premières pages trouvées correspondant à une requête donnée sont stockées pour former un premier corpus. Ce processus se repète pour autant de requêtes, c'est-à-dire, 869 fois.

À la fin de cette collecte, nous avons obtenu un peu plus de 17.000 documents[6] récupérés en 24 heures environ. Par la suite, nous avons fait subir certaines transformations à l'ensemble de ces corpus pour en enlever les marques correspondant au format HTML[7]. Pour ce faire nous avons utilisé un script *perl* de « déhtmlisation ».

Par exemple, voici un échantillon en HTML faisant partie de l'un des corpus obtenus à partir d'une requête :

```
<td><font face="Arial" size="5"><b>L'Ontario est-il &agrave; vendre ?</b>
</font>

<p><font face="Arial" size="2">R&eacute;ellement, l'Ontario <i>est-il</i>
&agrave; vendre ? Pour les personnes qui adorent l'Ontario et son
patrimoine naturel, et qui veulent qu'ils soient bien g&eacute;r&eacute;s,
c'est souvent la premi&egrave;re question que soul&egrave;ve l'Initiative
strat&eacute;gique concernant les terres (ISCT).</font></p>
```

Après transformations, cet échantillon dévient :

```
L'Ontario est-il à vendre ?

Réellement, l'Ontario est-il à vendre ? Pour les personnes qui adorent
l'Ontario et son patrimoine naturel, et qui veulent qu'ils soient bien
gérés, c'est souvent la première question que soulève l'Initiative
stratégique concernant les terres (ISCT).
```

Le résultat de l'ensemble des opérations de « déhtmlisation » donne un corpus global que nous appelons « corpus Web » qui contient 38.242.073 mots et 1.368.903 de phrases.

7.5.4 Extraction des patrons de cooccurrence

Pour l'obtention d'une base d'informations lexicales, c'est-à-dire des patrons de cooccurrence, nous avons analysé le corpus Web avec la grammaire G2 (le processus a duré environ 50 heures[8]). À la suite de cette analyse nous avons obtenu environ 4 millions de dépendances qui ont été transformées en patrons indépendamment de leur marquage (produites par une règle fiable ou moins fiable). En effet, nous avons considéré indifféremment les dépendances MF1 et MF2 car nous étions ici intéressée par la *quantité* de dépendances et non pas par leur *qualité*.

Le script de transformation des dépendances en patrons de cooccurrence permet l'obtention d'un format standard pour chaque rection (`X Prép`) et totalise la fréquence d'une rection donnée éliminant ainsi les doublons.

Il est à noter que, étant donné les erreurs lors de la *déhtmlisation* et lors de l'extraction de dépendances à partir d'un corpus bruité, le nombre de rections uniques extraites se

[6]Les tailles des corpus obtenus peuvent varier de façon significative car il peut arriver qu'il y ait moins de vingt pages trouvées pour une requête (les tailles des pages trouvées sont aussi très différentes).

[7]Au moment d'effectuer ce travail, ce n'est que le format texte qui sert d'entrée à l'analyseur.

[8]Nous avons utilisé une machine Pentium IV avec un processeur de 1,6 GHz.

situe autour de 100.000, mais un certain nombre d'entre elles sont erronés. Voici quelques exemples :

```
 1  identité de l
 1  impératif de &
 2  organisme de 1,75 \%
 2  \$ par mois
 2  \% de 2000
 2  2345678912526272829 30Taux de intérêt
 3  http://www.environment.detr.gov.uk/index.htm  sur développement
 3  ulfure dans roche
 4  nř95 de revue
 4  ouvrir à vendredi 5 février 1999
10  population de 25 avril
```

Ces erreurs n'affectent pas la suite de notre travail car l'algorithme de levée d'ambiguïtés ne compare que les rections extraites du corpus B, qui sont, par défaut, bien formés (voir 8.2).

7.5.5 Calcul des estimations des probabilités de rattachement

À partir des fréquences de chaque patron de cooccurrence ainsi que du calcul de la fréquence de chaque occurrence d'`X` dans les corpus extraits du Web lors des requêtes (et non pas dans l'absolu), nous avons calculé une mesure de la proportion de rattachement.

Nous appelons cette mesure *estimation de la probabilité* ou `EPR`[9] :

```
EPR(X,Prép) = fréq(X,Prép)/fréq(X)
```

La mesure `EPR` est calculée pour toutes les relations extraites du Web et stockée à côté du patron correspondant :

```
0.0043 corréler|par
0.0115 induire|plutôt que de
0.0002 coût|à titre de
0.1765 lemme|de
0.0027 immobilisation|dans le cadre de
0.0006 équivalent|avec
0.4167 enthalpie|de
0.0037 replacer|parmi
0.0011 apprécier|à la lumière de
0.0635 tremplin|pour
```

Cette mesure s'avère indispensable lors de la levée des ambiguïtés de rattachement, comme nous le montrons dans le chapitre suivant.

[9] [Vol01a] définit cette mesure comme le rapport entre la fréquence du bigramme *fréq(mot,préposition)* et la fréquence de l'unigramme *fréq(mot)*

7.6 Résumé

Dans ce chapitre, nous avons décrit les bases de la méthode que nous proposons pour le traitement efficace des ambiguïtés de rattachement prépositionnel dans le cadre d'un modèle d'analyseur syntaxique robuste. Cette méthode combine les résultats obtenus par des grammaires symboliques avec des techniques d'apprentissage.

Le point essentiel de notre approche réside dans l'introduction d'information lexicale, mais à la différence de l'information encodée dans des dictionnaires, cette information est pondérée avec une mesure statistique obtenue à partir d'un très grand corpus et concerne des patrons de cooccurrence variés (sans tenir compte de la distinction classique entre « argument », « complément », « modifieur », etc.). Ces patrons sont ainsi des cooccurences syntaxiques (et non pas des cooccurrences textuelles), c'est-à-dire des mots apparaissant ensemble dans une même relation de dépendance (attachements produits lors d'une première analyse). L'approche est non supervisée car les relations de dépendance sont produites par le parseur lors de cette première analyse.

Le travail décrit dans ce chapitre concerne principalement la création d'une grande base d'informations lexicales incorporée à la grammaire de l'analyseur. Nous présentons dans le chapitre suivant l'utilisation de cette base pour la levée des ambiguïtés de rattachement liées au rattachement prépositionnel.

Chapitre 8

Levée d'ambiguïtés liées au rattachement prépositionnel

8.1 Introduction

Dans le chapitre précédent nous avons présenté les lignes générales de la méthode que nous avons développée dans le but d'améliorer les résultats obtenus par le parseur lors de la désamguïsation du rattachement prépositionnel.

En effet, nous avons caractérisé les étapes principales de cette méthode et nous en avons décrit les phases initiales : un développement de grammaires à base de règles pour une première analyse du texte en entrée, suivi de l'utilisation des dépendances obtenues lors de cette analyse pour la création d'une base de données lexicales. Cette base, composée de patrons de sous-catégorisation pondérés, est fondamentale pour la levée des ambiguïtés de rattachement telle que nous la proposons.

Les paramètres initiaux étant présentés, ce chapitre décrit la deuxième étape de la méthode, c'est-à-dire l'algorithme de levée des ambiguïtés (section 8.2) et présente les résultats de l'évaluation de cette approche. Nous montrons aussi (section 8.3) différentes expériences menées avec des variations dans les paramètres utilisés pour la construction de la table de sous-catégorisation et nous comparons les résultats obtenus avec les résultats de la méthode principale.

La section finale (8.4) récapitule les conclusions dégagées de notre approche.

8.2 Levée d'ambiguïtés de rattachement

La méthode que nous avons développée pour la levée d'ambiguïtés de rattachement combine des informations lexicales (patron de cooccurrence d'une dépendance extraite par l'analyseur) avec des informations de nature statistique (poids d'un patron de co-occurrence, c'est-à-dire, fréquence du patron `X,Prép` par rapport à l'élément `X` de la dépendance).

L'algorithme de levée d'ambiguïtés de rattachement tient compte de ces informations au moment de trouver le bon rattachement pour une tête syntaxique donnée. En effet, lors d'une première analyse, le parseur applique une stratégie non-déterministe et produit

ainsi un certain nombre de relations de dépendance possibles entre un élément à rattacher (`N`) et une tête (`X`). Ces dépendances possibles sont calculées par différentes heuristiques encodées dans la grammaire à base de règles.

L'objectif de la levée d'ambiguïtés est de trouver le seul rattachement correct dans un contexte (phrase) donné.

Pour ce faire, à partir du texte analysé une première fois, les paramètres suivants sont utilisés :

- l'ensemble de dépendances extraites par des règles « moins fiables » (P < 93 %, *cf.* 7.4) du corpus à valider ;
- la table d'informations lexicales avec une mesure `EPR` (estimation de la probabilité de rattachement, *cf.* 7.5.5) pour chaque patron de cooccurrence.

En effet, le corpus à valider est analysé une première fois par l'analyseur et un ensemble de relations de dépendance sont produites. Les dépendances liées au rattachement prépositionnel présentent un trait indiquant une mesure de fiabilité attribuée selon la règle de grammaire qui les a produites (*cf.* 7.4). Pour rappel, le trait `MF1` correspond à des relations calculées par des règles fiables et le trait `MF2` par des règles moins fiables, la notion de fiabilité étant définie par un taux de précision supérieur ou inférieur à 93 %.

Nous considérons que les dépendances marquées avec `MF1` sont déjà correctes, c'est pourquoi, lors de la levée d'ambiguïtés, seulement les dépendances `MF2` sont prises en compte.

L'algorithme prend cet ensemble de dépendances à valider et compare les patrons de sous-catégorisation en conflit pour un même élément à rattacher, avec les patrons et leurs poids encodés dans la table d'informations lexicales. Le résultat de cette comparaison aboutit à la proposition d'un seul rattachement et à la suppression des dépendances qui ne correspondent pas au patron choisi (*cf.* 8.2.2).

8.2.1 Algorithme

Plus précisement, l'algorithme utilisé pour la levée d'ambiguïtés est constitué des actions suivantes :

1. Acquisition des données de la base lexicale : patrons de cooccurrence et mesures d'estimation de la probabilité de rattachement (`EPR`) pour chaque patron ;
2. Pour chaque phrase du fichier[1] :
   a) traiter les dépendances à désambiguïser :
      - parcours de toutes les dépendances `A_MF2(X,Prép,N)` pour une même phrase ;
      - identification d'un même `N` dans des dépendances avec `X Prép` différents ;
      - recherche de chaque `X Prép` dans la table de données lexicales.
      - si l'`EPR` existe pour un patron donné :
        - le comparer avec l'`EPR` du patron suivant ;
        - garder le patron avec l'`EPR` supérieur.

[1] Il est à noter que pour une phrase `P`, si l'analyseur a extrait une dépendance `A_MF1(X,Prép,N)`, il ne peut pas avoir extrait au même temps une dépendance `A_MF2(X',Prép',N)`, car les règles produisant des dépendances `MF2` ne s'appliquent *que* si un nom `N` n'a pas été rattaché par une règle `MF1`, *cf* 7.4.2.

- si un patron donné n'existe pas dans la table :
  - s'il y a un `EPR` supérieur pour le même `N`, alors le supprimer du fichier final ;
  - autrement le garder.

b) imprimer les dépendances `MF1` pour la phrase courante ;

c) imprimer les dépendances désambiguïsées ;

d) imprimer la phrase courante.

Deux cas de figure sont possibles lors de la comparaison de patrons de cooccurrence issus des dépendances à valider.

Les deux patrons à comparer existent

Lors du premier cas, les deux patrons (provenant de deux dépendances ayant le même élément `N` à rattacher) existent dans la base de patrons de cooccurrence.

Dans ce cas, leurs poids (mesures `EPR`) seront comparés et le patron avec le poids supérieur sera gardé.

Si une égalité des poids des patrons se produit, étant donné que l'algorithme ne dispose pas d'autres informations pour lever l'ambiguïté, les deux configurations sont alors gardées.

Un seul patron existe

Lors de l'existence d'un seul patron de cooccurrence, on suppose que le patron inexistant n'a pas été trouvé dans les corpus lors de la construction de la table. Son poids équivaut alors à zéro. La comparaison n'a pas lieu car c'est le patron existant dans la table qui est proposé.

Dans le cas où aucun des patrons n'existe dans la base, l'algorithme se retrouve devant un cas similaire à celui évoqué plus haut : à défaut d'autres informations pour lever l'ambiguïté, les deux patrons sont gardés.

8.2.2 Exemple d'application de l'algorithme

L'analyseur syntaxique extrait quatre relations de dépendance (liées au rattachement prépositionnel) à partir de la phrase suivante :

(1) *La méthode d'amortissement choisie aura un impact significatif sur les comptes.*

Parmi ces relations, il existe un rattachement fiable et un cas d'ambiguïté de rattachement manifesté par la présence de trois relations de dépendance pour un même élément à rattacher :

```
A_MF1(<méthode:1>,<de:2>,<amortissement:3>)
A_MF2(<avoir:5>,<sur:9>,<compte:11>)
A_MF2(<impact:7>,<sur:9>,<compte:11>)
A_MF2(<significatif:8>,<sur:9>,<compte:11>)
```

L'algorithme valide automatiquement la première dépendance et met en œuvre les comparaisons nécessaires entre les éléménts `X Prép` des dépendances avec le trait `MF2`.

La première comparaison concerne les patrons `avoir sur` et `impact sur`. La table de patrons de cooccurrences contient, parmi d'autres données :

```
...
0.0008 avoir|sans
0.0005 avoir|sous
0.0118 avoir|sur
...
0.0005 impact|selon
0.0626 impact|sur
...
0.0323 significatif|à
...
```

D'après les poids observés dans la base, c'est le deuxième patron (`impact sur`) qui est validé car son poids est le plus élévé :

```
0.0118 avoir|sur
0.0626 impact|sur
```

La dépendance `A(<avoir :5>,<sur :9>,<compte :11>)` est alors éliminée et la dépendance `A(<impact :7>,<sur :9>,<compte :11>)` potentiellement validée. Elle ne le sera pas définitivement tant qu'il restera pour la même phrase d'autres dépendances en concurrence (ayant le même élément `N`) à rattacher (dans l'exemple il reste la dépendance `A(<significatif :8>,<sur :9>,<compte :11>)`).

Par la suite, le patron `significatif sur` est recherché dans la table dans le but de comparer son `EPR` avec celui de `impact sur`. Or, le processus de comparaison se trouve ici dans le cas où l'un des patrons n'existe pas dans la table (c'est le cas pour `significatif sur`). Cela signifie qu'il n'a pas été attesté dans le corpus et donc que sa mesure `EPR` vaut zéro.

C'est donc le patron `impact sur` qui est validée et par conséquent la dépendance `A(<impact :7>,<sur :9>,<compte :11>)` est la seule gardée. La sortie définitive de l'analyse pour la phrase de l'exemple plus haut est donc :

```
A(<méthode:1>,<de:2>,<amortissement:3>)
A(<impact:7>,<sur:9>,<compte:11>)
```

Les marques de fiabilité ont disparu car on considère correctes toutes les relations produites[2].

[2]L'option de visualiser ou non ces marques se fait par cohérence avec les autres dépendances extraites par l'analyseur, dans le but d'obtenir une sortie homogène. Ces marques sont toujours accessibles car elles sont encodées sous forme de trait pour chaque dépendance liée au rattachement prépositionnel.

8.2.3 Évaluation

Nous avons évalué les résultats de l'ensemble de la méthode proposée sur le corpus B, utilisé pour valider aussi les gramaires initiales. Pour rappel, ce corpus est constitué de trois corpus des domaines économique, scientifique et juridique, extraits du Web, avec un total de 9.339 mots et 337 phrases (*cf.* Annexe E). Une version corrigée manuellement après une première analyse par le parseur atteste 1.089 dépendances liées au rattachement prépositionnel.

L'évaluation du résultat de la levée d'ambiguïtés est obtenu automatiquement grâce à un script qui compare la sortie de l'analyseur après cette étape de désambiguïsation avec le corpus validé manuellement. Le tableau 1 recapitule les résultats obtenus en utilisant la grammaire initiale et la grammaire de base, avec et sans levée d'ambiguïtés.

Grammaires initiales	Précision	Rappel	F1	Déps.
G1	87,49 %	77,04 %	**81,93 %**	959
G2	71,37 %	92,10 %	**80,40 %**	1413
G2 après désamb.	83,21 %	85,12 %	**84,16 %**	1120

Table 1. Résultats des grammaires G1 et G2 initiales et G2 après la levée d'ambiguïtés.

D'après ces résultats, il est facile de constater que la moyenne générale `F1` est supérieure dès lors qu'on combine la grammaire symbolique et l'apprentissage de la sous-catégorisation pour la levéee des ambiguïtés. Cette meilleure moyenne est obtenue par une baisse du taux de précision par rapport à la meilleure grammaire à base de règles (G1) mais par une augmentation du taux de rappel.

Il est important de signaler que, à la différence des résultats rapportés en début du chapitre précédent (section 7.2), ces résultats sont obtenus pour tout type d'ambiguïté de rattachement, c'est-à-dire, quelle que soit la configuration des constituants et non pas seulement pour la configuration classique `VP NP PP`. Des configurations avec plusieurs `PP`s sont ainsi prises en compte (`VP NP PP1 PP2 ... PPn`).

8.2.4 Analyse des erreurs

Globalement, les erreurs existantes sont de deux types. Elles peuvent provenir d'erreurs dans les étapes précédentes de l'analyse linguistique (segmentation, étiquetage, etc.) ou bien elles peuvent être dues aux informations erronées stockées dans la base de patrons de cooccurrence.

L'exemple suivant montre une erreur d'étiquetage au niveau de la relation de dépendance `A_MF2_5(<sommer :1>,<de :2>,<résultat :3>)` :

```
A_MF1_3(<solde:12>,<de:14>,<gestion:15>)
A_MF1_3(<résultat:3>,<de:4>,<exploitation:5>)
A_MF2_5(<sommer:1>,<de:2>,<résultat:3>)
A_MF2_7(<résultat:18>,<avant:20>,<impôt:21>)
A_MF2_9(<courant:19>,<avant:20>,<impôt:21>)
```

```
102>La somme du résultat d'exploitation et du résultat financier
fournit un solde intermédiaire de gestion : le résultat courant
avant impôts .
```

En effet, le nom *somme* a été etiqueté comme verbe par l'étiqueteur et une fausse relation (`<sommer :1>,<de :2>,<résultat :3>`) a été extraite. Étant donné qu'elle n'a pas dû être attestée dans la table de patrons, cette relation a été maintenue (et comptée comme erreur).

Les deux exemples suivants, en revanche, sont des erreurs dues aux informations erronées de la table (estimations des probabilités des patrons). Pour la phrase qui suit, trois relations sont en concurrence avec `<par :16>,<exploitation :18>` comme syntagme prépositionnel à rattacher :

```
A_MF1_3(<rentabilité:6>,<de:7>,<société:9>)
A_MF1_3(<effet:21>,<de:22>,<structure:24>)
A_MF2_5(<intégrer:11>,<par:16>,<exploitation:18>)
A_MF2_7(<excédent:14>,<par:16>,<exploitation:18>)
A_MF2_9(<dégagé:15>,<par:16>,<exploitation:18>)
104>Le résultat courant exprime clairement la rentabilité
de la société en intégrant à la fois les excédents dégagés
par l'exploitation et les effets de sa structure bilantielle.
```

Après désambiguïsation, on obtient la relation suivante (les deux autres proposées par l'analyseur initialement sont effacées) :

```
A_MF2_5(<intégrer:11>,<par:16>,<exploitation:18>)
```

En effet, dans la table de patrons de sous-catégorisation la valeur du patron `dégager par` (qui serait le patron correct pour cet exemple) est inférieure à celle de `intégrer par` :

```
0.0371 intégrer|par
...
0.0010 excédent|pour
...
0.0061 dégager|par
```

Dans ce cas, le corpus initial ne contient pas assez d'occurrences du patron `dégager par` pour le calcul des valeurs correctes des `EPR`. Le même problème se pose pour l'exemple suivant avec le syntagme `<dans :10>,<milieu :12>` à rattacher :

```
A_MF1_3(<propagation:5>,<de:6>,<énergie:8>)
A_MF2_5(<agir:2>,<de:3>,<propagation:5>)
```

```
A_MF2_5(<faire:20>,<dans:21>,<vide:23>)
A_MF2_7(<énergie:8>,<dans:10>,<milieu:12>)
A_MF2_8(<propagation:5>,<dans:10>,<milieu:12>)
A_MF2_9(<mécanique:9>,<dans:10>,<milieu:12>)
228>Il s' agit de la propagation d' une énergie mécanique dans un
milieu matériel : ce déplacement ne peut se faire dans le vide .
```

La relation proposée comme correcte est la suivante :

```
A_MF2_7(<énergie:8>,<dans:10>,<milieu:12>)
```

La table de patrons de sous-catégorisation contient les informations suivantes :

```
0.0206 énergie|dans
...
0.0117 propagation|dans
..
0.0003 mécanique|dans
```

Dans ce cas aussi, il n'y a pas assez d'occurrences dans les corpus collectés du patron `propagation dans`, qui serait la bonne solution pour cet exemple.

Malgré ces erreurs, les résultats globaux de la méthode démontrent l'avantage d'une approche hybride qui combine des grammaires symboliques avec des techniques d'apprentissage.

Nous allons par la suite montrer des variations réalisées autour des paramètres utilisés pour la mise en œuvre de la méthode telle qu'elle a été présentée jusqu'ici. Les résultats obtenus renforcent la conclusion dégagée à partir des chiffres présentés à la table 1.

8.3 Variations sur les paramètres

Nous avons voulu modifier certains paramètres utilisés lors de la réalisation de la méthode générale décrite jusqu'ici, dans le but de comparer les résultats et de valider notre approche.

Les variations sur les paramètres concernent :

1. l'acquisition de dépendances uniquement « fiables » provenantes de l'analyse du corpus B pour la construction de la table d'informations lexicales ;
2. la variation dans la taille du corpus à partir duquel on crée la table de patrons de sous-catégorisation ;
3. l'utilisation d'un seuil de fréquence (et non pas de la mesure `EPR`) pour la prise en compte des patrons de sous-catégorisation lors de la levée d'ambiguïités.

Nous présentons par la suite ces trois expériences ainsi que les résultats obtenus pour chacune d'elles.

8.3.1 Acquisition de dépendances uniquement « fiables »

Dans notre modèle, nous avons créé la table de patrons de cooccurrence à partir de toutes les dépendances extraites par l'analyseur lors d'une première analyse d'un très grand corpus (voir section 7.5).

Cependant, puisque notre grammaire permet de marquer les dépendances selon leur degré de fiabilité, nous avons fait l'hypothèse que l'utilisation *uniquement* des dépendances extraites par des règles fiables (`MF1`) permettra la création d'une table d'informations lexicales moins bruitée, c'est-à-dire avec des patrons de cooccurrence majoritairement corrects. On suppose alors, que cette table aidera à une meilleure levée d'ambiguïtés, ce qui permettra d'obtenir un meilleur taux de précision et une meilleure moyenne F1. La figure 8.1 schématise notre démarche.

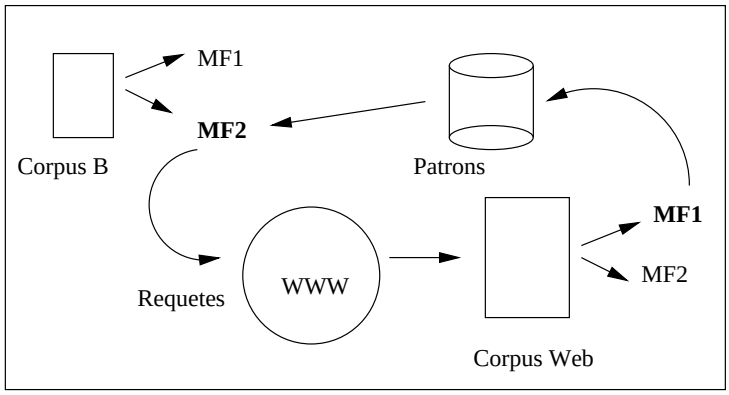

FIG. 8.1 – Apprentissage exogène avec uniquement des relations `MF1`.

En effet, on utilise les dépendances avec un taux de précision inférieur à 93 % (MF2) pour les transformer en requêtes pour le Web. Les résultats obtenus de ces requêtes constituent un corpus que nous analysons avec la grammaire. À la différence de l'expérience principale (schématisée par la figure 7.3 du chapitre précédent, où les patrons de cooccurrence sont extraits de l'ensemble de dépendances produites lors d'une première analyse) nous constituons ici la base d'exemples uniquement avec les patrons des dépendances « fiables », c'est-à-dire celles obtenues avec des règles ayant un taux de précision supérieur à 93 %.

Le processus de désambiguïsation se fait alors en utilisant la base d'exemples créée avec cet ensemble de relations « fiables ». La taille de cette table de patrons de cooccurrence est réduite environ de la moitié par rapport à l'expérience principale (elle contient encore quelques erreurs dues à la transformation initiale de dépendances en requêtes).

Comme pour l'expérience principale, des mesures `EPR` sont calculées pour chaque patron de cooccurrence (calcul de l'estimation de rattachement entre `X Prép` et `X`. Par la suite, l'algorithme de levée d'ambiguïtés utilise cette table pour le même corpus à valider.

Le tableau suivant montre les résultats de cette expérience :

Grammaires initiales	Précision	Rappel	F1
Toutes déps.	83,21 %	85,12 %	**84,16 %**
Uniquement déps. MF1	79,08 %	86,04 %	82,41 %

Il existe une amélioration par rapport à l'utilisation uniquement d'une grammaire à base de règles (rappel table 1, moyenne de 81,93 % pour la grammaire G1). En revanche, les résultats sont inférieurs par rapport à la méthode principale qui tient compte de l'ensemble des dépendances extraites lors de l'analyse du grand corpus, quel que soit leur mesure de fiabilité.

Ces résultats confirment que la quantité de dépendances à apprendre est plus important que le fait qu'elles soient bruitées : la quantité compense le bruit, ce qui conforte l'idée de l'utilisation d'un très grand corpus, le Web, même si une partie des données n'est pas « propre ».

8.3.2 Utilisation de corpus d'apprentissage plus petits

Pour la deuxième expérience, nous avons voulu créer la table de patrons de cooccurrence à partir de différents corpus, plus petits que le Web et des mêmes domaines que le corpus à valider (économique, scientifique, juridique) :

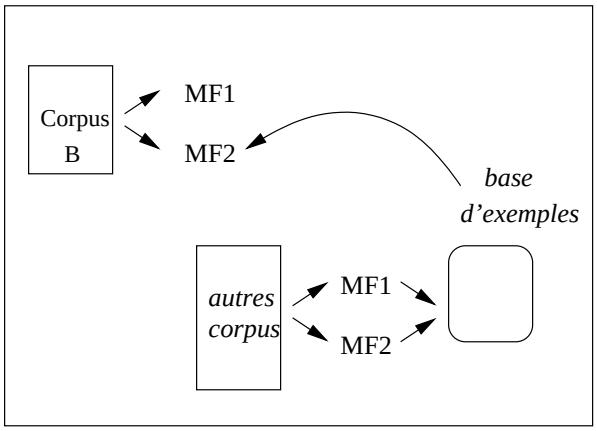

FIG. 8.2 – Apprentissage exogène avec des relations `MF1` et `MF2`.

L'ensemble du corpus provenant du Web représente 150 fois le corpus B initial : il a environ 38 millions de mots (38.242.073 mots). Nous avons fait l'hypothèse qu'un corpus plus petit (environ 30 fois supérieur au corpus à valider, c'est-à-dire 250.178 mots) mais plus propre (mêmes domaines et genres, sans erreurs dues à la déhtmlisation, etc.) pourrait favoriser la création d'une table de patrons de cooccurrence plus efficace pour la levée d'ambiguïtés de rattachement.

Pour cette expérience, ainsi que pour l'expérience principale (*cf.* figure 7.3) nous avons utilisé pour la création de la base d'exemples des corpus différents au corpus à valider.

Nous avons aussi fait une expérience d'apprentissage *endogène* : nous avons utilisé le même corpus à valider comme source pour l'extraction des patrons de cooccurrence (stratégie similaire à [Bou94]) :

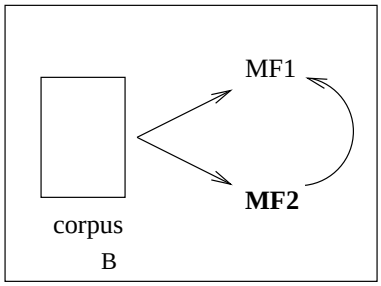

FIG. 8.3 – Apprentissage endogène avec uniquement des relations `MF1`.

Dans les deux cas, avec le corpus d'environ 250.000 mots et avec le même corpus à valider, la méthode a été la même : première analyse, obtention d'un ensemble de dépendances, transformation en patrons de cooccurrence et calcul de leurs mesures d'estimation de la probabilité de rattachement. Pour le grand corpus, la base contient environ 15.000 patrons ; pour le corpus à valider, environ 800.

L'algorithme de levée d'ambiguïtés a été appliqué en utilisant ces deux bases différentes. Les résultats sont exprimés dans la table suivante :

	Précision	Rappel	F1
G1	87,49 %	77,04 %	81,93 %
Corpus Web	83,21 %	85,12 %	**84,16 %**
Grand corpus	79,92 %	86,87 %	83,25 %
Corpus B	82,19 %	85,58 %	82,41 %

On constate que, même avec des corpus beaucoup plus petits que notre corpus tiré du Web à partir d'une première analyse du corpus B, les résultats sont meilleurs en combinant grammaire symbolique et apprentissage. Cette idée renforce la conclusion à laquelle nous sommes arrivés plus haut. En revanche, les différents taux de précision et aussi les moyennes générales sont légèrement inférieures à celles obtenues avec le très grand corpus. Cela montre une fois de plus, que l'important pour cette méthode est d'obtenir une base de patrons de cooccurrence la plus grande possible : la taille est donc plus significative que la « qualité » des données.

8.3.3 Calcul d'un seuil de fréquence

Nous avons voulu utiliser une mesure statistique différente de la mesure `EPR` comme information sur un patron de cooccurrence donné. Nous avons ainsi mesuré la fréquence attestée dans le corpus (Web) du patron de cooccurrence `X Prép`. Cette fréquence est

donc relative au corpus dont on extrait le patron et non au patron de cooccurrence dans l'absolu.

L'hypothèse faite au départ de cette expérience est la suivante : plus un patron est fréquent dans le corpus, plus il est fiable. Nous avons ainsi créé la table de patrons de cooccurrence par rapport à différents seuils établis pour la mesure de fréquence.

Une première table a été créée comme table de base autorisant tout patron, quel que soit son seuil de fréquence. Par la suite, d'autres tables avec des seuils supérieurs à 5, 10, 15, 20 et 25 ont été construites. Pour chaque table, l'algorithme de levée d'ambiguïtés a été appliqué sur l'ensemble de dépendances du même corpus à valider.

Le tableau suivant synthétise quelques uns des résultats obtenus, tout en les comparant à ceux obtenus en utilisant la mesure `EPR` :

	Précision	Rappel	F1
Mesure `EPR(X,Prép)`	**83,21 %**	85,12 %	**84,16 %**
fréquence (X,Prép) > 0	81,60 %	83,29 %	82,43 %
fréquence (X,Prép) > 5	80,89 %	83,38 %	82,12 %
fréquence (X,Prép) > 10	80,05 %	83,47 %	81,73 %
fréquence (X,Prép) > 25	78,56 %	83,93 %	81,16 %

En effet, on constate que l'hypothèse faite en début de cette expérience n'est pas validée : l'augmentation de la fréquence des seuils est plutôt négative pour la levée d'ambiguïtés liées au rattachment prépositionnel. Cela s'explique facilement, car plus le seuil est élevé, plus le nombre de patrons de cooccurrence diminue et, comme nous l'avons demontré plus haut, l'important est d'avoir une base de patrons la plus riche possible.

Voici le nombre de patrons obtenus pour chacun des seuils établis :

Seuil de fréquence	Nombre de patrons
Fréquence (X,Prép) = 5	6432
Fréquence (X,Prép) = 10	2109
Fréquence (X,Prép) = 15	940
Fréquence (X,Prép) = 20	615
Fréquence (X,Prép) = 25	449

Une fois de plus, la quantité de données s'avère un paramètre plus significatif que tout autre type de critère, en l'occurence la fréquence d'apparition, plus haut (8.3.1 et 8.3.2) la « propreté » des données.

8.4 Résumé

Dans cette troisième partie de la thèse, constituée des chapitres 7 et 8, nous avons décrit une méthode permettant l'amélioration du résultat des analyses pour le rattachement prépositionnel. Cette technique, telle que nous l'avons présentée au chapitre précédent, est fondée sur une division de l'analyse en deux étapes : extraction de relations de dépendance pour les transformer en patrons de cooccurrence intégrés dans la grammaire (avec

des mesures statistiques) et utilisation de ces patrons pour la levée d'ambiguïtés liées au rattachement prépositionnel.

Le résultat d'une telle approche implique une lexicalisation de la grammaire initiale, à laquelle on aboutit par la combinaison de la grammaire symbolique avec des techniques d'apprentissage.

Comme nous le montrons dans ce chapitre, les meilleurs résultats (basés sur une moyenne du taux de précision et de rappel, `F1` = 84,16 %) sont obtenus avec un grand corpus (le Web) pour la création de la base d'informations lexicales (nombre plus élévé de patrons) et une mesure de calcul de l'information mutuelle entre les éléments intégrant ces patrons.

Notre approche permet ainsi d'améliorer les résultats obtenus uniquement par une grammaire à base de règles. Les informations lexicales et statistiques obtenues sur un très grand corpus s'avèrent très utiles pour une meilleure résolution des ambiguïtés structurelles liées au rattachement prépositionnel.

Conclusion

Contexte

L'analyse syntaxique robuste (*robust parsing*) est devenue une technique essentielle pour toute application qui touche au contenu des documents (recherche d'information, systèmes de question-réponse, etc.). De par leur définition, les analyseurs robustes sont amenés à traiter des grandes quantités de corpus hétérogènes et à toujours produire une analyse réutilisable dans d'autres applications.

Cependant, une des difficultés majeures à laquelle se heurtent ces systèmes est la production d'analyses avec un taux de précision élevé et stable quelles que soient les caractéristiques du corpus en entrée (domaine « standard » ou spécialisé, données « propres » ou bruitées, encodage enrichi ou appauvri, etc.).

Ainsi, il existe des phénomènes présents dans tous les types de corpus, généralement bien modélisés dans la plupart des systèmes. Pour ces phénomènes, les stratégies adoptées par les analyseurs existants donnent des résultats significatifs (par exemple, plus de 90 % de taux de précision pour le marquage de syntagmes nominaux, pour l'extraction de la relation de dépendance sujet[3] etc.).

En revanche, il existe aussi un ensemble de phénomènes plus « complexes » du point de vue de leur traitement automatique, qui de ce fait sont mal modélisés dans les grammaires des analyseurs (ou pas modélisés du tout). D'une part, il s'agit de phénomènes considérés au-delà de la linguistique, requérant une prise en compte spécifique qui met en jeu des paramètres liés à la ponctuation et à la visualisation de la structure du document. D'autre part, il s'agit de phénomènes linguistiques dont une analyse précise dépasse le niveau strictement syntaxique.

Bilan

À partir de ces constats, nous avons proposé et implémenté une architecture pour un analyseur robuste capable de faire face à ces problèmes, c'est-à-dire capable de traiter du texte libre appartenant à différents domaines avec une couverture et une précision élevées et homogènes.

Plus précisément, nous avons conçu un modèle d'analyseur syntaxique qui maîtrise la variété des phénomènes linguistiques et structurels présents dans des corpus tout venant et garantit la qualité des analyses produites.

[3] Pour la relation sujet-verbe, la précision est de 92,5 % pour IFSP, l'analyseur du français de XRCE [AMC97a], et de 95 % pour l'analyseur de l'anglais de l'université de Helsinki [TJ97].

Pour ce faire, dans un premier temps, notre travail a consisté à caractériser et à évaluer quelques systèmes existants dans le but de mieux cerner d'une part leurs caractéristiques et, d'autre part, d'identifier précisément les phénomènes qui font baisser leurs performances.

Cela nous a permis d'envisager une architecture basée sur la modularité et l'adaptabilité des grammaires : reconfiguration automatique de différents modules grammaticaux selon le type de phénomène présent dans l'entrée à analyser et acquisition automatique d'informations lexicales, puisées dans de grandes collections de données, dans le but d'enrichir les grammaires de dépendances en vue d'améliorer la résolution de certains cas d'ambiguïté structurelle.

Nous avons mis en œuvre cette approche sur la base de la plate-forme XIP [AMCR02] tout en profitant des avantages que ce formalisme présente (souplesse pour l'enrichissement des règles, facilité d'ajouter ou d'enlever des modules grammaticaux, possibilité d'articuler plusieurs ressources, etc.).

Par rapport à l'analyseur XIP-F existant, la nouvelle architecture que nous avons définie est basée sur une analyse syntaxique à deux niveaux : selon les caractéristiques structurelles de chaque unité en entrée, différentes grammaires sont appliquées, tout d'abord une grammaire noyau (pour les phrases appelées `N1`), ensuite des grammaires spécialisées au traitement de différents phénomènes (pour les phrases `N2`). Préalablement à l'application de la grammaire noyau, un module de prétraitement identifie des unités au-delà de la définition classique des phrases (listes et titres).

L'application de l'ensemble de ces grammaires que nous avons construites garantit un taux de précision constant pour environ deux tiers des corpus analysés quel que soit leur domaine. En effet, la grammaire noyau *couvre* environ 20 - 25 % des phrases de tout corpus et le pourcentage de phrases *couvertes* par les grammaires spécialisées est d'environ 17 % pour la ponctuation (phénomènes de clôture –guillemets et parenthèses–, 15 % pour les titres et de 10 % pour listes et énumérations). La notion de « couverture » doit être comprise comme le pourcentage de phrases analysées pour lesquelles le parseur est capable de donner un taux de précision élevé (entre 90 % et 96 %) pour le *chunking* et pour l'extraction de la dépendance de base sujet.

Cette notion de « couverture » est liée à la notion de « fiabilité » de l'analyse : environ deux tiers des phrases des corpus de domaines variés sont analysés avec un taux de précision assez haut ; le tiers de phrases restant est aussi analysé, mais la précision de cette analyse reste moins fiable (c'est-à-dire que des segments et des dépendances sont produites mais leur précision risque d'être plus faible à cause de la structure et des phénomènes particuliers dans la phrase jusqu'ici non modélisés). Notre approche modulaire permet d'améliorer les performances linguistiques de l'analyseur sans pour autant nuire à sa couverture.

À côté de cette approche modulaire, nous avons proposé une méthodologie d'apprentissage non supervisé[4] permettant l'enrichissement de l'analyseur avec un lexique de patrons de cooccurrence pondérés, obtenus automatiquement à partir de grandes collec-

[4]L'approche est non supervisée car les relations de dépendance sont produites automatiquement par l'analyseur.

tions de données dans le Web. Concrètement, les poids donnés à ces patrons sont des poids calculés sur des cooccurrences syntaxiques et non pas des cooccurrences purement textuelles : ils sont associés à des mots apparaissant initialement dans une même relation de dépendance (des attachements produits par l'analyseur lors d'une première analyse), ce qui permet d'éliminer certains cas où un lien serait impossible (par exemple, entre deux propositions différentes).

Pour ce faire, ces dépendances sont transformées en requêtes soumises à un moteur de recherche sur le Web (sous la forme « `X NEAR Prép` », où `X` est un nom, un verbe ou un adjectif et `Prép` une préposition). Les documents obtenus sont alors analysés et les dépendances liées au rattachement prépositionnel sont toutes stockées avec, pour chacune, une mesure d'estimation de la probabilité de rattachement calculée sur la base de la fréquence d'apparition dans les corpus collectés. Lors de la désambiguïsation entre deux dépendances, l'algorithme choisit celle ayant un poids plus élevé ; en cas de poids égaux, c'est l'heuristique du rattachement le plus proche qui est appliquée.

Dans l'ensemble, cette technique permet d'améliorer la précision des dépendances liées au rattachement prépositionnel quelle que soit sa configuration : 71,37 % de précision initiale avec l'analyseur à base de règles, 83,21 % après désambiguïsation en utilisant le lexique acquis automatiquement avec des informations de sous-catégorisation pondérées (la moyenne F1 est de 80,40 % initialement et de 84,16 % après l'application de notre méthode).

Perspectives

La grammaire noyau et les grammaires spécialisées produisent une analyse fine pour environ 65 % des phrases des corpus tout venant. Des évaluations de la dépendance `sujet` montrent qu'il n'y a pas de variation en qualité selon les différents phénomènes. Nous obtenons une analyse globalement fiable dont la qualité reste stable à travers l'hétérogénéité des corpus.

Cependant, quelques phrases présentent encore des problèmes, par exemple, celles contenant des phénomènes non modélisés (des propositions interrogatives, exclamatives, impératives) :

“ Quel rôle veut-on donner aux salariés ?”

“ La preuve, la kit-car et la T4 l'utilisent ! ”

“ Soulignons en outre que les centrales brûlent quelque 44 millions de tonnes de charbon dur chaque année.”

L'ensemble des phrases de ce type ne représente qu'entre 5 % et 7 % dans nos corpus (certainement plus dans des corpus de type manuel d'instructions) ; une grammaire spécialisée permetrait un traitement spécifique de ces phénomènes (nous ne l'avons pas modélisée et implémentée dans le cadre de notre travail faute de temps).

En revanche, il existe d'autres phénomènes pour lesquels des solutions diverses, plus ou moins complexes, sont envisageables.

a) Premièrement, il serait souhaitable d'améliorer certaines règles des grammaires existantes, dans le but d'augmenter la précision et le rappel (cas non ou mal repérés alors

qu'ils sont, *a priori*, modélisés). De plus, il faudrait étendre le nombre de règles pour prendre en compte des sous-cas non modélisés jusqu'ici (par exemple, listes où l'amorce n'a pas deux points, listes imbriquées, titres avec ponctuation finales). Exemples :

"La vaporisation de l'essence est une étape importante dans le processus de carburation, et plusieurs facteurs sont à prendre en compte.

1) La pression qui, à une température donnée, tend un liquide à s'évaporer d'autant plus vite que la pression est basse. (la pression atmosphérique est plus basse à haute altitude qu'au niveau de la mer)

2) La température, qui sous la pression fait évaporer les liquides plus ou moins rapidement."

Cet exemple n'est pas repéré comme liste parce que l'amorce ne finit pas par deux points.

"Les Etats contractants accorderont aux réfugiés résidant régulièrement sur leur territoire le même traitement qu'aux nationaux en ce qui concerne les matières suivantes :

a) Dans la mesure où ces questions sont réglementées par la législation ou dépendent des autorités administratives : la rémunération, y compris les allocations familiales lorsque ces allocations font partie de la rémunération, la durée du travail, les heures supplémentaires, (...) ;

b) La sécurité sociale (les dispositions légales relatives aux accidents du travail, aux maladies professionnelles, à la maternité, à la maladie, à l'invalidité, à la vieillesse et au décès, au chômage, aux charges de famille, ainsi qu'à tout autre risque qui, conformément à la législation nationale, est couvert par un système de sécurité sociale), sous réserve :

i) Des arrangements appropriés visant le maintien des droits acquis et des droits en cours d'acquisition ;

ii) Des dispositions particulières prescrites par la législation nationale du pays de résidence (...)."

Quant à cet exemple, le problème réside dans l'identification de la liste imbriquée.

"Les bougies : il est conseillé de vérifier leur état régulièrement."

Ici, le titre n'est pas identifié à cause de la ponctuation.

b) Une autre source importante de problèmes qui influent sur les résultats de l'analyseur sont les erreurs liées à des étapes de traitement linguistique précédent (normalisation, segmentation ou étiquetage morphologique). Pour ce problème, il faudrait améliorer les ressources et les outils de traitement morphologique.

c) Enfin, il existe encore un certain nombre de phénomènes requérant une analyse plus poussée souvent au-delà de la syntaxe (par exemple les coordinations ambiguës).

"Cela réduit la valeur de l' immeuble d'autant au bilan et au bout de vingt ans cet immeuble figurera donc pour une valeur nulle ."

Ainsi, dans cet exemple, une analyse superficielle donnerait les deux syntagmes prépo-

sitionnels avec la préposition *au* comme coordonnés, alors qu'ici la coordination se situe à un niveau supérieur (il s'agit des deux propositions).

En ce qui concerne la méthode proposée pour améliorer les résultats liés à la désambiguïsation du rattachement prépositionnel, elle peut être étendue dans le but de résoudre des cas jusqu'à présent problématiques (mauvais calcul des mesures d'estimation des probabilités de rattachement).

Une extension de cette méthode consisterait à utiliser les trois éléments d'une relation de dépendance (`X Prép Y`) pour le calcul de la mesure d'estimation de la probabilité de rattachement et non pas celle de son patron de sous-catégorisation (constitué des deux premiers éléments). Cette technique demanderait la construction d'une base de *trigrammes* à partir d'un corpus encore plus grand pour qu'il soit suffisamment représentatif de la présence des trois éléments.

Au niveau de l'algorithme, en cas d'ambiguïté, la priorité serait donnée à la structure constituée des trois éléments, mais au cas où elle ne serait pas attestée dans la table, la recherche serait effectuée par rapport aux patrons de sous-catégorisation.

Une autre option à explorer est celle de la généralisation de la méthode avec l'utilisation de classes sémantiques. Ces classes seraient intégrées sous forme de traits pour chaque élément de la relation, et l'algorithme utiliserait ce critère (concordance de classe sémantique entre une tête et un élément à rattacher) pour effectuer la levée de l'ambiguïté.

Finalement, il est envisageable que cette méthode puisse être adaptée à d'autres cas de problèmes d'ambiguïté structurelle, notamment le cas de la coordination.

Bibliographie

[AB99] A. Abeillé et P. Blache. Grammaires et analyseurs syntaxiques. *Traité IC2, volume Ingénierie des Langues*, 1999.

[Abn90] S. Abney. Rapid incremental parsing with repair. Dans *Proceedings of the 6th New OED Conference*, pages 1–9, University of Waterloo, Ontario, 1990.

[Abn91] S. Abney. Parsing by chunks. Dans *Principle-Based Parsing*, édité par R. Berwick, S. Abney, et C. Tenny. Academic Publishers, 1991.

[Abn94] S. Abney. Partial Parsing, 1994. Tutorial given at ANCL-94, http ://www.sfs.nphil.uni-tuebingen.de/~abney/Papers.html.

[Abn95] S. Abney. Chunks and dependencies : bringing processing evidence to bear on syntax. Dans *Linguistics and Computation*, édité par J. Cole, G. Green, et J. L. Morgan, pages 145–164. CSLI Publications, Stanford, 1995.

[Abn96] S. Abney. Chunk stylebook, 1996. Non publié, http ://www.vinartus.net/spa/publications.html.

[AJJ+93] D. Appelt, Hobbs J., Bear J., D. Israel, et M. Tyson. 'FASTUS' : a finite-state processor for information extraction from real-world text. Dans *Proceedings of IJCAI-93*, Chambéry, 1993.

[AMC97a] S. Aït-Mokhtar et J. P. Chanod. Incremental Finite-State Parsing. Dans *Proceedings of the 8th Conference on Applied Natural Language Processing, ANLP-97*, pages 72–79, Washington, 1997.

[AMC97b] S. Aït-Mokhtar et J. P. Chanod. Subject and Object Dependency Extraction Using Finite-State Transducers. Rapport technique, Rank Xerox Research Centre (RXRC), Grenoble, 1997.

[AMCR01] S. Aït-Mokhtar, J. P. Chanod, et C. Roux. A multi-input dependency parser. Dans *Proceedings of the Internationl Workshop on Parsing Technologies, IWPT-01*, pages 201–204, Beijing, Chine, 2001.

[AMCR02] S. Aït-Mokhtar, J. P. Chanod, et C. Roux. Robustness beyond shallowness : Incremental deep parsing. Dans *Special Issue of the Natural Language Engineering Journal on Robust Methods in Analysis of Natural Language Data*, pages 121–144. Cambridge University Press, 2002.

[AMGP98] S. Aït-Mokhtar et N. Gala Pavia. A description of the English Finite-State Parser Architecture. Rapport technique, Xerox Research Centre Europe (XRCE), Grenoble, 1998.

[AMP+99] G. Adda, J. Mariani, P. Paroubek, M. Rajman, et J. Lecomte. L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Cahiers/Langues. Évaluation des systèmes de traitement automatique*, 2(2), 1999.

[ASS96] S. Abney, R. E. Schapire, et Y. Singer. Boosting applied to tagging and pp attachment. Dans *Proceedings of Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pages 38–45, Maryland, 1996.

[ASU88] A. V. Aho, R. Sethi, et J. D. Ullman. Syntax analysis. Dans *Compilers. Principles, techniques and tools*, pages 159–278. Addison-Wesley Publishing Company, Stanford, 1988.

[BF00] D. Bourigault et C. Fabre. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, 25 :139–151, 2000.

[Bib93] D. Biber. Using register-diversified corpora for general language studies. Dans *Computational Linguistics 19(2)*, pages 219–241, 1993.

[Blo33] L. Bloomfield. *Language*. Holt, (second edition 1961), New York, 1933.

[BM97] C. Berthouzoz et P. Merlo. Statistical Ambiguity Resolution for Grammar-based Parsing. Dans *Current Issues in Linguistic Theory. Selected papers from Recent Advances in Natural Language Processing, RANLP'97*, édité par N. Nicolov et R. Mitkov. John Benjamins, Amsterdam, 1997.

[BN81] C. Boitet et N. Nédobejkine. Recent developments in Russian-French Machine Translation at Grenoble. *Linguistics*, 19 :199–271, 1981.

[Bou94] D. Bourigault. *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes.* Thèse de doctorat, École d'Hautes Études en Sciences Sociales, Paris, 1994.

[BP92] R. Basili et M. T. Pazienza. A shallow syntactic analyzer to extract word associations from corpora. Dans *Literary and Linguistics Computing*, pages 114–124, Cargese, Corse, 1992.

[BPL99] A. Ballim, V. Pallotta, et Ch. Lieske. Robust Text Analysis : an Overview. Rapport technique, Swiss Federal Institute of Technology, Lausanne, 1999.

[BPV94] R. Basili, M. T. Pazienza, et P. Velardi. A "not-so-shallow" parser for collocational analysis. Dans *15th International Conference on Computational Linguistics, COLING-94*, Kyoto, Japon, 1994.

[BPZ99] R. Basili, M. T. Pazienza, et F. M. Zanzotto. Lexicalizing a shallow parser. Dans *Proceedings of TALN'99*, Cargese, Corse, 1999.

[BR94] E. Brill et P Resnik. A rule-based approach to prepositionnal phrase attachment disambiguation. Dans *15th International Conference on Computational Linguistics, COLING-94*, pages 1198–1204, Kyoto, Japon, 1994.

[Bri94] T. Briscoe. Parsing (with) Punctuation etc. Rapport technique, Rank Xerox Research Centre (RXRC), Grenoble, 1994.

[Bru98] C. Brun. *Étude et implantation de la coordination en vue de l'analyse automatique du français écrit dans le cadre de la Grammaire Lexicale Fonctionnelle.* Thèse de doctorat, Université Grenoble III, 1998.

[BSA98] M. Bayraktar, B. Say, et V. Akman. An analysis of English punctuation : the special case of comma. *International Journal of Corpus Linguistics*, 3(1) :33–57, 1998.

[Bus00] C. Bush. Analyse des déclencheurs des énumérations d'entités nommées sur le Web. Rapport technique, LIMSI numéro 2000-05, Orsay, 2000.

[CB95] M. Collins et J. Brooks. Prepositional phrase attachment through a backed-off model. Dans *Proceedings of the third Workshop on Very Large Corpora*, Cambridge, Massachussets, 1995.

[Cha94] J. P. Chanod. Développements en Analyse Syntaxique Automatique. Dans *Proceedings of TALN-94*, Marseille, 1994.

[Cha00] J. P. Chanod. Robust Parsing and Beyond. Dans *Robustness in Language Technology*, édité par G. Van Noord et JC Junqua. Kluwer, 2000.

[Chu88] K. Church. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143, 1988.

[CT96] J. P. Chanod et P. Tapanainen. A Robust Finite-State Parser for French. Dans *ESSLII'96 Robust Parsing Workshop*, Prague, 1996.

[Dai94] B. Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* Thèse de doctorat, Université de Paris 7, 1994.

[DBV99] W. Daelemans, S. Buchholz, et J. Veenstra. Memory-based shallow parsing. Dans *Proceedings of CoNLL-99*, pages 53–60, Bergen, 1999.

[DDC97] J. Dubois et F. Dubois-Charlier. *Dictionnaire des verbes français.* Larousse, Paris, 1997.

[DDTS99] L. Dini, V. Di Tomaso, et F. Segond. Ginger II, an example-driven word sense disambiguator. *Computers and Humanities, special issue*, 1999.

[Deb82] F. Debili. *Analyse Syntaxico-Sémantique Fondée sur une Acquisition Automatique de Relations Lexicales-Sémantiques.* Thèse de doctorat, Université Paris XI, 1982.

[DH96] S. Douglas et M. Hurst. Layout and Language : lists and tables in technical documents. Dans *Proceedings of ACL-SIGPARSE*, pages 19–24, Santa Cruz, California, 1996.

[DSCH97] D. Denis, A. Sancier-Chateau, et M. Huchon. *Encyclopédie de la Grammaire et de l'Ortographe.* Librairie Grénérale Française, Paris, 1997.

[Dub69] J. Dubois. *Grammaire structurale du français.* Larousse, Paris, 1969.

[Eje93] E. Ejerhed. Nouveaux courants en analyse syntaxique. *T.A.L.*, 34(1), 1993.

[FF02] C. Fabre et C. Frérot. Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus. Dans *Actes de la conférence TALN-02*, Nancy, 2002.

[FGHP+01] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, et C. Jacquemin. Document selection refinement based on linguistic features for QALC, a question

answering system. Dans *Proceedings of Recent Advances in Natural Language Processing, RANLP-2001*, Tzigov Chark, Bulgarie, 2001.

[Fil68] C. J. Fillmore. The Case for Case. Dans *Universals in Linguistic Theory*, édité par E. Bach et R.T. Harms, pages 1–88. Holt, Rinehart and Winston, New York, 1968.

[FLM99] M. Finkelstein-Landau et E. Morin. Extracting semantic relationships between terms : Supervised vs. unsupervised methods. Dans *Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure*, pages 71–80, Dagstuhl Castle, Germany, 1999.

[Fra78] L. Frazier. *On comprehending sentences : syntactic parsing strategies.* Thèse de doctorat, University of Connecticut, 1978.

[GC01] E. Gaussier et N. Cancedda. Probabilistic models for PP-attachment resolution and NP analysis. Dans *Proceedings of ACL-2001, Computational Natural Language Learnings Workshop, CoNLL-2001*, pages 45–52, Toulouse, 2001.

[GKPS85] G. Gazdar, E. Klein, G. Pullum, et I. Sag. *Generalized Phrase Structure Grammar.* Blackwell, Harvard University Press, Cambridge, MA, 1985.

[GP99] N. Gala Pavia. Using the incremental finite-state architecture to create a Spanish shallow parser. Dans *Proceedings of XV Congres of SEPLN*, pages 75–82, Lleida, Spain, 1999.

[GP00] N. Gala Pavia. Hétérogénéité des corpus : vers un parseur robuste reconfigurable et adaptable. Dans *RECITAL-00 (session étudiante de TALN)*, pages 477–482, Lausanne, 2000.

[GP01] N. Gala Pavia. A two-tier corpus-based approach to robust syntactic annotation of unrestricted corpora. *T.A.L.*, 42(2) :381–411, 2001.

[Gre92] G. Grefenstette. SEXTANT : Exploring unexplored contexts for semantic extraction from syntactic analysis. Dans *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, ACL-92*, Newark, Delaware, 1992.

[Gre93] M. Grevisse. *Le bon usage. Grammaire Française.* Duculot, Paris, 1993.

[Gre94] G. Grefenstette. SEXTANT : Extracting semantics from raw text, implementation details. Dans *Integrated Computer-Aided Engineering , Vol. 6, No. 4 (Special Issue on Knowledge Extraction from Text).* Wiley and Sons, New York, 1994.

[Gre95] G. Grefenstette. Low-level parsing applied to technical manuals. Dans *Proceedings of IPSM'95, Industrial parsing of Software Manuals*, Limerick, Ireland, 1995.

[Gre99] G. Grefenstette. The World Wide Web as a resource for example-based machine translation tasks. Dans *Proceedings of Aslib Conference on Translating and the Computer 21*, London, 1999.

[Gre02] G. Grefenstette. Natural Language Processing at the limit between symbolic and numeric processing. Dans *Conférence invitée au 13ème congrès francophone de Reconnaissance des Formes et Intelligence Artificielle, RFIA-02*, Angers, 2002.

[Gsi93] Gsi-Erli, France. *Le dictionnaire AlethDic*, 1.5 edition, 1993.

[GV97a] E. Giguet et J. Vergne. From part-of-speech tagging to memory-based deep syntactic analysis. Dans *Proceedings of the International Workshop on Parsing Technologies, IWPT-97*, pages 77–88, Boston, MA, 1997.

[GV97b] E. Giguet et J. Vergne. Syntactic analysis of unrestricted French. Dans *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP-97*, pages 276–281, Tzigov Chark, Bulgaria, 1997.

[Hal94] M. A. K. Halliday. *Introduction to Functional Grammar*. Edward Arnold, second edition, London, 1994.

[Hea92] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. Dans *Proceedings of the Fourteenth International Conference on Computational Linguistics, COLING-92*, pages 539–545, Nantes, France, 1992.

[Hin83] D. Hindle. User manual for Fidditch, a deterministic parser. *Naval Research Laboratory Technical Memorandum 7590-142*, 1983.

[HR93] D. Hindle et M. Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1) :103–120, 1993.

[Hud94] R. A. Hudson. Word grammar. Dans *The Encyclopedia of Language and Linguistics*, édité par R. Asher. Pergamon Press, 1994.

[HZ80] K. Hamill et A. Zamora. The use of titles for automatic document classification. *American Society for Information Science*, pages 396–402, 1980.

[Ill00] G. Illouz. *Typage de données textuelles et adaptation des traitements linguistiques. Application à l'annotation morpho-syntaxique*. Thèse de doctorat, Université de Paris-Sud, UFR Scientifique d'Orsay, 2000.

[Jac90] Paul S. Jacobs. To parse or not to parse : Relation-driven text skimming. Dans *13th International Conference on Computational Linguistics, COLING-90*, pages 194–198, Helsinki, 1990.

[Jac97] C. Jacquemin. Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus, 1997. Habilitation à diriger des recherches, Institut de recherche en informatique de Nantes.

[JB87] K. Jensen et J.-L. Binot. Disambiguating prepositionnal phrase attachments by using on-line dictionnary definitions. *Computational Linguistics*, 13(3-4) :251–260, 1987.

[JB00] C. Jacquemin et C. Bush. Combining lexical and formatting cues for named entity acquisition from the Web. Dans *Proceedings of Joint Sigdat Conference On Empirical Methods In Natural Language Processing And Very Large Corpora (EMNLP/VLC-2000)*, édité par H. Schutze, Hong Kong, 2000.

[JBR02] B. Jacquemin, C. Brun, et C. Roux. Semantic enrichment for information extraction using word sense disambiguation. Dans *Workshop on creating and using semantics for information retrieval and filtering, International conference on Language Ressources and Evaluation, LREC-02*, 2002.

[Jen92] K. Jensen. PEG : The PLNLP English Grammar. Dans *Natural language processing : the PLNLP approach*, édité par K. Jensen, G. Heidorn, et S. Richardson, Boston, 1992. Kluwer Academic Publishers.

[Jes69] O. Jespersen. *La syntaxe analytique*. Eds. de Minuit, Paris, 1969.

[JLT75] A. K. Joshi, L. Levy, et M. Takahashi. Tree adjunct grammars. *Journal of the Computer and System Sciences*, 10(1) :136–163, 1975.

[Jon94] B. E. M. Jones. Exploring the role of punctuation in parsing natural text. Dans *15th International Conference on Computational Linguistics, COLING-94*, pages 421–425, Kyoto, Japan, 1994.

[Jon96] B. Jones. *What's the Point ? A (Computational) Theory of Punctuation.* Thèse de doctorat, Centre for Cognitive Science, University of Edimburgh, 1996.

[Jos85] A. K. Joshi. How much context-sensitivity is necessary for characterizing structural descriptions –Tree Adjoining Grammars. *Natural Language Processing –Theoretical, Computational and Psychological Perspectives*, 1985.

[Jos96] A. Joshi. A parser from antiquity : An early application of finite-state transducers to natural language parsing. Dans *Proceedings ECAI'96, workshop on extended finite state models of language*, Budapest, 1996.

[JR94] C. Jacquemin et J. Royauté. Retrieving terms and their variants in a lexicalized unification-based framework. Dans *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 132–141, Dublin, 1994. Springer Verlag.

[Kah00] S. Kahane. Les grammaires de dépendance. *Traitement Automatique des langues*, 42(1), 2000.

[KB82] R. Kaplan et J. Bresnan. The mental representation of grammatical relations. Dans *Lexical-Functional Grammar : A Formal System for Grammatical Representation*, pages 173–381. Bresnan, J., Cambridge, MA, 1982.

[Kim73] J. Kimball. Seven principles of surface structure parsing in natural language. *Cognition*, 2 :15–47, 1973.

[KVHA95] F. Karlsson, A. Voutilainen, J. Heikkilä, et A Anttila. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin and New York, 1995.

[Kwo75] K. L. Kwok. *The use of titles and cited titles as document representation for automatic classification.* Information Processing and Management 11 (8/12), 1975.

[Lal97] J. Lallot. Apollonius dyscolus [de constructione] (de la construction = peri syntaxeos / apollonius dyscole). Texte grec accompagné de notes critiques,

introduction, traduction et notes exégétiques. Paris : Vrin. 2 v. (303 ; 476 pages), 1997.

[Leb02] T. Lebarbé. *Hiérarchie inclusive des unités linguistiques en analyse syntaxique coopérative.* Thèse de doctorat, Université de Caen, 2002.

[LGDM+99] C. Luc, C. Garcia-Debanc, M. Mojahid, M.-P. Péry-Woodley, et J. Virbel. A linguistic approach to some parameters of layout : a study of enumerations. Dans *Proceedings of AAAI Fall Symposium on Using Layout for the Generation, Understanding or Retrieval of Documents*, North Falmouth, MA, 1999.

[LM98] A. López et M. Montes. Nominalization in titles : a way to extract document details. Dans *Memoria del Simposium Internacional de Computación CIC-98*, Mexico, 1998.

[Luc00] Ch. Luc. *Représentation et composition de structures visuelles et rhétoriques du texte.* Thèse de doctorat, Université de Toulouse, 2000.

[Man01] F. Maniez. L'ambiguité syntaxique due aux structures coordonnées en anglais médical : analyse de la performance d'un logiciel d'aide à la traduction. Dans *RECITAL-01 (session étudiante de TALN)*, Tours, 2001.

[Mar85] A. Martinet. *Syntaxe Générale.* Armand Colin, Paris, 1985.

[MCB97] P. Merlo, M. W. Crocker, et C. Berthouzoz. Attaching multiple prepositionnal phrases : generalized backed-off estimation. Dans *Proceedings of the second conference on Empirical Methods in Natural Language Processing, EMNLP-97*, 1997.

[Mel88] I. A. Mel'cuk. *Dependency Syntax : theory and practice.* Albany, The SUNY Press, New York, 1988.

[Mey87] C.F. Meyer. *A linguistic study of American puntuation.* Peter Lang, 1987.

[MHJ81] L. A. Miller, G. E. Heidorn, et K. Jensen. Text-critiquing with the EPISTLE system : an author's aid to better syntax. Dans *Proceedings of the National Computer Conference*, pages 649–655, 1981.

[Mil80] L. A. Miller. Project EPISTLE : A system for the automatic analysis of business correspondence. Dans *Proceedings of the first annual national Conference on Artificial Intelligence*, pages 280–282, 1980.

[Mon02] L. Monceaux. *Adaptation du niveau d'analyse des interventions dans un dialogue.* Thèse de doctorat, Université de Paris-Sud, UFR scientifique d'Orsay, 2002.

[Mor99] E. Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques.* Thèse de doctorat, Université de Nantes, 1999.

[MPRZ99] M. Munoz, V. Punyakanok, D. Roth, et D. Zimak. A learning approach to shallow parsing. Dans *Proceedings EMNLP-WVL'99*, pages 168–178, 1999.

[NBH01] G. Nunberg, T. Briscoe, et R. Huddleston. Punctuation. Dans *Cambridge Grammar of English*, édité par R. Huddleston et G. Pullum, 2001.

[Nun90] G. Nunberg. *The Linguistics of Punctuation.* CSLI Lecture Notes, CSLI Publications, Stanford, CA, 1990.

[Per83] D. M. Perlmutter. *Studies in Relational Grammar 1.* University of Chicago Press, Chicago, 1983.

[PRL00] D. Proux, F. Rechenmann, et Julliard L. A pragmatic information extraction strategy for gathering data on genetic interactions. Dans *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'2000)*, pages 279–285, La-Jolla, California, August 19-23, 2000.

[Pro01] D. Proux. *Muninn : une stratégie d'extraction d'informations dans des corpus spécialisés par application de méthodes d'analyse linguistique de surface et de représentation conceptuelle des structures sémantiques.* Thèse de doctorat, Université de Bourgogne, 2001.

[PS87] C. Pollard et I. Sag. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago, 1987.

[PS96] R. Pasero et P. Sabatier. GNF : une grammaire noyau du français. Rapport technique, Laboratoire d'Informatique Fondamentale de Marseille (LIM), Marseille, 1996.

[Roc93] E. Roche. *Analyse syntaxique transformationnelle du français par des transducteurs et lexique-grammaire.* Thèse de doctorat, Université Paris 7, 1993.

[Rou96] C. Roux. *Une méthode efficace de parsage des grammaires syntagmatiques généralisées.* Thèse de doctorat, Université de Montréal, 1996.

[RRR94] A. Ratnaparkhi, J. Reynar, et S. Roukos. A maximum entropy model for prepositional phrase attachment. Dans *ARPA Workshop on Human Language Technology*, Plainsboro, 1994.

[RRR97] A. Ratnaparkhi, J. Reynar, et S. Roukos. A maximum entropy model for prepositional phrase attachment. Dans *Proceedings of the 5th Conference on Applied Natural Language Processing, ANLP-97*, pages 16–19, Washington DC, 1997.

[SA97] B. Say et V. Akman. Current Approaches to Punctuation in Computational Linguistics. *Computers and the Humanities*, 30(6) :457–469, 1997.

[SA98] B. Say et V. Akman. Dashes as typographical cues for the information structure. Dans *Proceedings of 3rd Conference in Information-Theoretic approaches to logic, language and computation (ITALLC '98)*, pages 209–223, Hsitou, Taiwan, 1998.

[SBB+87] J. Slocum, W. S. Bennet, J. Bear, M. Morgan, et R. Root. METAL : The LRC machine translation system. Dans *Machine Translation today : the state of the art*, édité par S. Michaelson et Y. Wilks, pages 319–350. Edinburgh University Press, 1987.

[Sch95] C. T. Schutze. PP-attachment and argumenthood. Rapport technique, MIT Working papers in Linguistics, 1995.

[Sch96] A. Schiller. Multilingual finite-state NP extraction. Dans *Proceedings of the ECAI 96 Workshop on Extended FS Models of Language*, 1996.

[SN97] J. Stetina et M. Nagao. Corpus based PP attachment ambiguity resolution with a semantic dictionary. Dans *Proceedings of the 5th Workshop on Very Large Corpora, VLC-97*, édité par J. Zhou et K. Church, pages 66–80, Beijing and Hongkong, 1997.

[Sri93] B. Srinivas. Punctuation and parsing of real-world texts. Dans *Twente Workshop on Language Technologies*, pages 163–167, Twente, 1993.

[Sri97] B. Srinivas. *Complexity of lexical descriptions and its relevance to partial parsing.* Thèse de doctorat, University of Pennsylvania, 1997.

[ST93] D. Sleator et D. Temperley. Parsing English with a Link Grammar. Dans *Proceedings of the 3rd International Workshop on Parsing Technologies, IWPT-93'*, 1993.

[Tap99] P. Tapanainen. *Parsing in two frameworks : finite-state and functional dependency grammar.* Thèse de doctorat, University of Helsinki, Language Technology, Department of General Linguistics, Faculty of Arts, 1999.

[Tes59] L. Tesnière. *Elements de syntaxe structurale.* Hachette (3e éd. 1979), Paris, 1959.

[TJ97] P. Tapanainen et T. Jarvinen. A non-projective dependency parser. Dans *Conference on Applied Natural Language Processing, ANLP-97*, pages 64–71, Washington, 1997.

[Tro97] F. Trouilleux. Identification et classement automatique des noms propres dans des textes en français. Dans *Mémoire de DEA*, GRIL, Université Blaise-Pascal, Clermont Ferrand, 1997.

[Tro01] F. Trouilleux. *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français.* Thèse de doctorat, GRIL, Université Blaise-Pascal, Clermont Ferrand, 2001.

[Vau71] B. Vauquois. Modèles pour la traduction automatique. Dans *Mathématiques et Sciences Humaines*, numéro 34, pages 61–71, Paris, 1971.

[Vau73] B. Vauquois. Les sytèmes informatiques et l'analyse de texte. Dans *L'analyse des corpus linguistiques. Problèmes et méthodes de l'indexation maximale.*, Strasburg, 1973.

[Ver02] J. Vergne. Une méthode pour l'analyse descendante et calculatoire de corpus multilingues : application au calcul des relations sujet-verbe. Dans *Actes de TALN-2002*, Nancy, 2002.

[Vir89] J. Virbel. Linguistic knowledge an text structures. Dans *Structured Documents.* André, J. and Quint, V. and Furuta, R., Cambridge University Press, 1989.

[VJ96] A. Voutilainen et T. Jarvinen. Using the English Constraint Grammar parser to analyse a software manual corpus. Dans *Industrial Parsing of Software Manuals.* Language and Computers : Studies in practical linguistics, 17, Rodopi, Amsterdam, 1996.

[Vol00] M. Volk. Scaling up. Using the WWW to resolve PP attachment ambiguities. Dans *Proceedings of Konvens*, pages 151–155, Ilmenau, 2000.

[Vol01a] M. Volk. The automatic resolution of prepositional phrase attachment ambiguities in German. Dans *Habilitationschrift*, Université de Zurich, 2001.

[Vol01b] M. Volk. Exploiting the WWW as a corpus to resolve PP attachment. Dans *Proceedings of Conference on Corpus Linguistics*, pages 601–606, Lancaster, 2001.

[Weh92] E. Wehrli. The IPS system. Dans *14th International Conference on Computational Linguistics, COLING-92*, pages 870–875, Nantes, 1992.

[WF96] H. Wu et T. Furugori. A hybrid disambiguation model for prepositional phrase attachment. *Literary and Linguistic Computing*, 11(4) :187–192, 1996.

[Whi95] M. White. Presenting punctuation. Dans *Proceedings from European Workshop on Natural Language Generation*, pages 107–125, Leiden, Pays Bas, 1995.

[ZDV97] J. Zavrel, W. Daelemans, et J. Veenstra. A rule-based approach to prepositionnal phrase attachment disambiguation. Dans *Conference on Computational Natural Language Learning, CoNLL-97*, pages 136–144, Madrid, 1997.

Annexes

Annexe A

Corpus pour l'étude initiale

A.1 Description

Le corpus utilisé pour les expériences initiales (repérage de phénomènes problématiques du point de vue de leur traitement automatique[1], développement de la grammaire noyau et des grammaires spécialisées) est constitué d'un ensemble de corpus collectés sur le Web (courant 2000) appartenant à des domaines et genres différents :

- juridique (rapport d'audiences, projets de loi),
- scientifique (médecine, physique, télécommunications),
- économique (rapport financier, journal *Les Echos*),
- journalistique (*Le Monde*, *Libération*, *L'Humanité*)
- technique (manuel d'instructions, description véhicule).

L'ensemble de ces corpus en français contient 153.023 mots et 5.205 phrases (notion de phrase élargie avec titres et listes, comptage effectué après identification de ces unités). La moyenne globale est d'environ 29,4 mots par phrase.

Le tableau suivant résume les caractéristiques de ces corpus.

Domaine		*Nbre de mots*	*Nbre de phrases*	*Longueur moyenne*
Juridique	23,1 %	35.339	1.005	35,16 mots
Scientifique	20,2 %	30.908	906	34,11 mots
Économique	20 %	30.664	1.111	27,60 mots
Journalistique	20,3 %	31.028	1.148	27,03 mots
Technique	16,4 %	25.084	1.037	24,19 mots
total	100 %	153.023	5.207	29,4 mots

[1]Pour cette tâche permettant d'évaluer les sorties des analyseurs IFSP [AMC97a] et SEXTANT [Gre92] un corpus en anglais a aussi été utilisé. Il provient de la base de données du centre de recherche de Xerox et est composé de textes provenant du *Wall Street Journal* ainsi que d'un corpus de biologie.

A.2 Échantillon des corpus

Les fragments qui suivent appartiennent aux différents corpus en anglais et français. Les corpus en français présentent les marques fournies par les pré-traitements (décrits au chapitre 5.2).

Corpus économique

<tit> Services traiteurs , services de restauration et services institutionnels </tit>

<list> L' Afrique du Sud compte plus de 25 000 restaurants et services annexes ayant un chiffre d'affaires combiné annuel d' environ 8,5 milliards de rands <sp>.</sp> Les deux segments visés sont <sp> :</sp> commercial (hôtels , restaurants , restauration rapide , stations de vacances) <sp>.</sp> institutionnel (hôpitaux , établissements carcéraux , maisons d' enseignement , forces armées , sociétés minières) <sp>.</sp> </list>

Trois entreprises dominent le segment des institutions et des services de cafétérias pour employés , et toutes ont des liens directs avec des fabricants de produits alimentaires .

L' approvisionnement de ces entreprises est assuré principalement par les fabricants et certains grossistes .

Les hôtels sont d' importants acheteurs de produits alimentaires sur le marché des restaurants et services annexes .

Ce secteur est en plein essor et , avec les prévisions fixant le nombre de touristes à un million par année d' ici 1996 , il devrait représenter des débouchés de plus en plus intéressants .

Comme c' est le cas dans les autres segments du marché , des liaisons importantes existent entre les établissements hôteliers et les conglomérats .

Les restaurants et établissements de restauration rapide forment 29 % du marché des restaurants et services annexes et 76 % du segment commercial .

La restauration rapide à la nord-américaine n' est pas encore ancrée dans les moeurs sud-africaines , mais elle devrait connaître un croissance considérable grâce à l' urbanisation accrue et à la hausse du revenu disponible prévues prochainement .

Les chaînes McDonald et Burger King songent à s' installer sur ce marché .

<tit> 2 <sp>.</sp> 4 Politique agricole et alimentaire </tit>

Avant les élections de 1994 , le Congrès National d' Afrique (ANC) , a élaboré un document intitulé " Plan de reconstruction et de développement " (RDP) qui a servi de plateforme électorale à ce parti pour faire connaître ses objectifs .

Sous le gouvernement majoritaire , dirigé par l' ANC , le Plan a été adopté comme étant un " document provisoire " du gouvernement de l' unité nationale (GNU) .

(...)

Corpus technique

<tit> Allumecigare </tit>

Un allumecigare peut être monté au tableau de bord à la place de l' obturateur .

Pour utiliser l' allumecigare , enfoncer le bouton au maximum .

Il ressort automatiquement lorsque l' élément électrique a atteint la température correcte .

<tit> (Accessoire en option) </tit>

<tit> Commutateur d' éclairage accessoires </tit>

Un commutateur peut être monté à cet emplacement pour commander d'autres accessoires électriques (voir page 53) .

<tit> Commandes de chauffage </tit>

Levier supérieur (7) à droite - air dirigé vers le plancher de la cabine .

Levier supérieur (7) à gauche - air dirigé vers les aérateurs audessus du tableau de bord .

Déplacer le levier inférieur (8) vers la droite pour augmenter la température de l' air fourni par le chauffage .

Déplacer le levier vers la gauche pour obtenir de l' air non chauffé aux ouïes d' aération .

<tit> Commande d' inclinaison du volant de direction </tit>

Pour incliner le volant , tourner le levier de serrage dans le sens contraire des aiguilles d' une montre pour débloquer le volant .

Le resserrer lorsque le volant est à l' angle voulu .

(...)

Corpus scientifique

<tit> Résultats </tit>

Le mélanome est une tumeur rare dont l' incidence est en forte augmentation dans les pays industrialisés : doublement de l' incidence en dix à quinze ans .

En France , on peut estimer l' incidence actuelle du mélanome entre 5000 et 6000 cas par an .

Deux types de facteurs de risque ont été identifiés dans le mélanome : des facteurs individuels , constitutifs (phototypes , nombreux nævus , nævus atypiques , antécédents personnels ou familiaux de mélanome) et un facteur comportemental , l' abus d' exposition au soleil ou aux ultraviolets (UV) artificiels (niveau de preuve A) .

Le geste diagnostique , sauf exception , doit être l' exérèse complète de la lésion cutanée suspecte (standard) .

L' ensemble de la pièce doit être confié à l' anatomopathologiste (standard) .

Le compte rendu anatomopathologique doit , au minimum , comporter le diagnostic de la nature mélanocytaire et la malignité , l' épaisseur maximum en millimètres (selon la méthode de Breslow) , l' état des marges d' exérèse , le niveau d' invasion (niveau de Clark) , l' existence ou non d' une zone de régression et l' existence ou non d' une ulcération .

(...)

Plus de 11800 niveaux et de 3450 raies intenses , incluant de nombreuses données nouvelles , ont été rapportées dans un ouvrage publié en 1992 [1].

L' analyse de ces spectres n' est pas achevée pour autant ; la recherche des niveaux de U IV , de Es I et II , l' interprétation de très nombreux niveaux excités et de leurs structures hyper fines sont autant de raisons de poursuivre l' étude spectroscopique des plus grands édifices atomiques connus .

Le besoin de données spectroscopiques précises pour répondre à la haute qualité des spectres pris sur le Hubble Space Telescope a poussé à réviser et étendre les analyses de Ce III , Pr III , Nd II , Tb II , Er III [2] , Au II et Au III .

La classification de deux autres spectres a débuté ici , Pt III en 1992 et Dy III [3] en 1994 , dont les raies ont été respectivement reconnues dans les étoiles chimiquement particulières c Lupi et HR465 .

(...)

Corpus juridique

Il en résulte que les écoles publiques font partie intégrante du code complet de l' art. 93 .

Par conséquent , les écoles publiques sont implicitement mais néanmoins clairement visées par le régime établi à l' art. 93 et sont donc protégées contre toute contestation fondée sur la Constitution ou la Charte .

Cette protection existe même si les droits relatifs aux écoles publiques ne sont pas eux-mêmes constitutionnalisés .

C' est le pouvoir absolu de la province de légiférer relativement aux écoles publiques , qui sont accessibles à tous les membres de la société , sansdistinction , qui est constitutionnalisé .

La province demeure libre d' exercer , comme elle l' entend , son pouvoir absolu en matière d' éducation , sous réserve des restrictions relativesaux écoles séparées imposées par le par. 93 (1) .

Le pouvoir législatif de la province n' est pas limité aux régimes d' écoles publiques et d' écoles catholiques romaines .

Cependant , une loi relative à l' éducation pourrait être assujettie à un examen fondé sur la Charte chaque fois que le gouvernement décide d' aller au-delà de son mandat de financer les écoles catholiques romaines séparées et les écoles publiques .

Les juges Sopinka et Major : Rien dans le par. 93 (3) de la Loi constitutionnelle de 1867 ne restreint le financement d'autres écoles .

Il prévoit explicitement l' exercice du pouvoir absolu pour établir un ń système d ' écoles séparées ou dissidentes ż là où il n' en existe pas .

Ni la loi sur le financement des écoles publiques ni une loi prescrivant le financement des écoles des appelants ne seraient visées par le par. 93(3) .

Une telle loi ne porterait pas atteinte à un droit ou à un privilège de la minorité protestante ou catholique romaine et elle n' établirait pas un système d ' écoles séparées ou dissidentes .

Lorsque la province exerce son pouvoir absolu en dehors des domaines mentionnés aux par. 93 (1) et (3) , une distinction qui viole la Charte n' est pas « expressément autorisée » ni même prévue .

Dans ces cas , la situation de la loi en cause n' est pas différente de celle d' une loi adoptée en vertu de l'un ou l' autre des chefs de compétence énumérés à l' art. 92 .

Appliquer la Charte n' invalidera aucun pouvoir conféré par l' art. 93 .

Parce qu' ils sont pertinents pour déterminer les droits et privilèges des écoles catholiques romaines , certains droits et privilèges des écoles protestantes de la majorité ne bénéficient pas , par le fait même , de la protection de la Constitution .

Les droits et privilèges des écoles publiques ne constituent que des points de repère pour déterminer les droits et privilèges des écoles séparées .

(...)

Corpus journalistique

<tit> « Un Paris traditionnel à droite et un plus moderniste à gauche » </tit>

Recueilli par CHRISTOPHE ALIX pour Liberation .

<tit> Le lundi 19 mars 2001 </tit>

Bruno Gouyette analyse les changements et les permanences de la géographie politique parisienne à l'issue d' un scrutin qui a vu le basculement de la capitale à gauche pour la première fois .

La gauche victorieuse à Paris , c' est un bouleversement de la carte électorale parisienne ?

La victoire de Bertrand Delanoë n' est pas un séisme électoral en faveur de la gauche : le rapport droite - gauche est de 50,4 % pour la droite et 49,6 % pour la gauche .

La gauche a eu 5.000 voix de moins que la droite ! Quant au rapport en élus , la gauche plurielle ne dispose que de 56 % des sièges du conseil de Paris .

Ce n' est donc pas un raz-de-marée qui aurait bouleversé la donne politique à Paris .

S' il y a une nouveauté radicale dans l' élection du 19 mars , c' est dans l' élection d' un maire qui se consacre uniquement et totalement à Paris.

C' est un vrai changement par rapport à la stratégie de Jacques Chirac de 1977 à 1995 .

Il n' y aurait donc pas de changement radical dans la nouvelle carte électorale issue du scrutin ? L' élection de dimanche n' est pas vraiment révolutionnaire pour qui se souvient des résultats des élections municipales de 1977 -les « élections matricielles » à Paris , car ce sont les premières municipales du genre dans la capitale .

En fait , les mêmes arrondissements avaient voté à gauche ou étaient en passe de le faire en mars 1977 : les IIe , IIIe , XIe , XIIIe , XVIIIe , XIXe et XXe avec une majorité à gauche et les Xe , XIIe et XIVe très près de passer à gauche .

Il y a vingt-quatre ans , un même écart de quelques milliers de voix séparait la droite et la gauche .

Ce qui était exceptionnel en fait , c' étaient les élections de 1983 et de 1989 , qui avaient donné tous les arrondissements à la même majorité de droite .

(...)

Corpus économique en anglais

Here are some changes in Social Security benefits and taxes that take effect with the new year :

BENEFITS : Monthly benefit checks increase 4.7 percent to offset the effects of inflation. The average retired worker's Social Security check will rise from $541 to $566. The increase also applies to recipients of Supplemental Security Income.

PAYROLL TAX : The tax rate rises from 7.51 percent to 7.65 percent. Employees will pay $765 for each $10,000 of earningssubject to tax.

TAX ON SELF-EMPLOYED : The rate rises to 15.30 percent. In 1989, the rate technically was 15.02 percent but the self-employed got an automatic 2 percent tax credit, so they paid 13.02 percent. They lose the 2 percent credit in 1990, but get to deduct half of Social Security taxes as a business expense.

WAGE BASE : Workers and self-employed individuals will pay Social Security taxes on the first $51,300 of earnings in 1990, up from $48,000 in 1989.

MEDICARE PREMIUM : Elderly and disabled workers will be charged $28.60 a month in 1990 for supplementary Medicare coverage, down from $31.90 in 1989. The decrease reflects the repeal of catastrophic health-care benefits. Medicare premiums are deducted directly from Social Security checks.

EARNINGS LIMIT : Social Security will allow beneficiaries ages 65 through 69 to earn $9,360 a year before reducing their benefits, up from $8,880 in 1989.

(...)

Corpus scientifique en anglais

P0- wg is required for allocation and for initiation of Ba expression that marks the allocation process. P0- Competence of cells to express wg is independent of their ability to receive the hh signal. P0- Wild type wg alleles transfected into Drosophila tissue culture cells display wg protein on the cell surface and in the extracellular surface, whereas mutant proteins appear not to be secreted. P0- Mosaic analysis suggests that wg-expressing cells sustain en expression only in adjoining cells. P0- wg is a target gene of the homeotic selector genes and is regulated by exd. P0- wg and dpp are required for target field neurons to adopt their proper fates and to send axons into the developing target structure. P0- Germ line transformation demonstrates that the 3' untranslated region of wg is sufficient for apical localisation of the wg transcript. P0- Mitotic recombination clones homozygous for deficiencies of wts show overgrowth and abnormal morphogenesis indicating that wts is a tumor suppressor gene. P0- Mutations at the y locus during IR hybrid dysgenesis involve integrating not resident I-element. P0- Immunohistochemical methods were used to study the temporal and spatial expression patterns of the y gene product in embryonic and pupal development to elucidate y polypeptide function. P0- coli bind specifically to the fat body enhancer (FBE) of Yp1 and Yp2. P0- Deletion analysis of the Yp1 upstream sequences has demonstrated that additional sequences are functionally equivalent to the fat body enhancer (FBE) in Yp1 upstream regionP0- z interacts with w in an eye specific manner. P0- zen is required for the normal ontogeny of the zen pattern and fating of the amnioserosa. P0- Mutations in zygotic dorsal class gene zen do not interact with RpII140 [wimp].

Annexe B

Traitements linguistiques dans l'analyseur XIP-F

Nous montrons dans cet annexe les différents traitements linguistiques fournis par les composants de l'analyseur XIP-F pour la phrase suivante :

La hausse des prix de l'aluminium réduira les marges des fabricants de boîtes .

B.1 Analyse morphologique

Les sorties correspondant à l'analyse morphologique sont produites par le module NTM (*Normalisation, Tokenisation et Morphological analysis*). Le fragment suivant correspond à la segmentation (*tokenisation*) et à l'étiquetage (*tagging*) des différents items de la phrase en exemple :

```
La      le      +Fem+SG+Def+Det+DET_SG
La      le      +Acc+Fem+SG+P3+PC
La      la      +Masc+InvPL+Noun+NOUN_INV

hausse  hausser +avoir+se+SN+IndP+SG+P3+Verb+VERB_P3SG
hausse  hausser +avoir+se+SN+IndP+SG+P1+Verb+VERB_P1P2
hausse  hausser +avoir+se+SN+Imp+SG+P2+Verb+VERB_P1P2
hausse  hausser +avoir+se+SN+SubjP+SG+P3+Verb+VERB_P3SG
hausse  hausser +avoir+se+SN+SubjP+SG+P1+Verb+VERB_P1P2
hausse  hausse  +deSN+Fem+SG+Noun+NOUN_SG

des     de      +Prep=le+InvGen+PL+Def+Prep+PREP_DE
des     un      +InvGen+PL+Indef+Det+DET_PL

prix    prix    +deSN+Masc+InvPL+Noun+NOUN_INV

de      de      +InvGen+PL+Indef+Det+DET_PL
```

```
de      de      +Prep+PREP_DE

l'      le      +Acc+InvGen+SG+P3+PC
l'      le      +InvGen+SG+Def+Det+DET_SG

aluminium       aluminium       +Masc+SG+Noun+NOUN_SG

réduira réduire +se+enSN+SN+aSVINF+avoir+aSN+deSN+Fut+SG+P3+Verb+VERB_P3SG

les     le      +Acc+InvGen+PL+P3+PC
les     le      +InvGen+PL+Def+Det+DET_PL

marges  marger  +SN+avoir+SubjP+SG+P2+Verb+VERB_P1P2
marges  marger  +SN+avoir+IndP+SG+P2+Verb+VERB_P1P2
marges  marge   +deSN+Fem+PL+Noun+NOUN_PL

des     de      +Prep=le+InvGen+PL+Def+Prep+PREP_DE
des     un      +InvGen+PL+Indef+Det+DET_PL

fabricants      fabricant       +deSN+Masc+PL+Noun+NOUN_PL

de      de      +InvGen+PL+Indef+Det+DET_PL
de      de      +Prep+PREP_DE

boîtes  boîter  +IndP+SG+P2+Verb+VERB_P1P2
boîtes  boîter  +SubjP+SG+P2+Verb+VERB_P1P2
boîtes  boîte   +Fem+PL+Noun+NOUN_PL

.       .       +SENT

\n\n    CR      +SENT
```

B.2 Analyse syntaxique : découpage en syntagmes

Liste de syntagmes et segments potentiellement fournis par XIP-F :

- `GROUPE` (phrase entière) ;
- `SC` (*sentence clause*), proposition ;
- `BG` (*begin group*), début proposition subordonnée ;
- `NP` (*noun phrase*), syntagme nominal ;
- `PP` (*prepositional phrase*), syntagme prépositionnel ;
- `AP` (*adjective phrase*), syntagme adjectif ;
- `FV` (*finite verb*), syntagme verbal conjugué ;
- `IV` (*infinitive verb*), syntagme verbal infinitif ;

– `GV` (*gerund verb*), syntagme verbal participe présent.

Sorties de l'analyseur XIP-F pour la phrase en exemple :

(La hausse des prix de l'aluminium réduira les marges des fabricants de boîtes .)

1. Affichage avec le texte en sortie parenthésé et des informations morphologiques complètes :

```
0>GROUPE(0:14){SC(0:7){NP(0:1){DET{La^le^+Fem+SG+Def+Det+DET_SG:0},
NOUN{hausse^hausse^+deSN+Fem+SG+Noun+NOUN_SG:1}},PP(2:3){PREP{des^
de^+Prep=le+InvGen+PL+Def+Prep+PREP_DE:2},NP(3:3){NOUN{prix^prix^+
deSN+Masc+InvPL+Noun+NOUN_INV:3}}},PP(4:6){PREP{de^de^+Prep+PREP_DE:
4},NP(5:6){DET{l'^le^+InvGen+SG+Def+Det+DET_SG:5},NOUN{aluminium^
aluminium^+Masc+SG+Noun+NOUN_SG:6}}},FV(7:7){VERB{réduira^réduire^
+se+enSN+SN+aSVINF+avoir+aSN+deSN+Fut+SG+P3+Verb+VERB_P3SG:7}}},
NP(8:9){DET{les^le^+InvGen+PL+Def+Det+DET_PL:8},NOUN{marges^marge^
+deSN+Fem+PL+Noun+NOUN_PL:9}},PP(10:11){PREP{des^de^+Prep=le+InvGen+
PL+Def+Prep+PREP_DE:10},NP(11:11){NOUN{fabricants^fabricant^+deSN+
Masc+PL+Noun+NOUN_PL:11}}},PP(12:13){PREP{de^de^+Prep+PREP_DE:12},
NP(13:13){NOUN{boîtes^boîte^+Fem+PL+Noun+NOUN_PL:13}}},SENT{.^.^+
SENT:14}}
```

Dans cette sortie :

– les syntagmes montrent les numéros du premier au dernier item qui les compose :

  `{PP(12:13)}`

– chaque unité lexicale présente la catégorie, la forme de surface, le lemme (séparés par ^), les traits morphologiques associés (séparés par +) et le numéro de position (séparé par :).

  `{NOUN{boîtes^boîte^+Fem+PL+Noun+NOUN_PL:13}`

2. Affichage simplifié tout en conservant quelques informations morphologiques :

```
0>GROUPE(0:14){SC(0:7){NP(0:1){DET{La^le:0},NOUN{hausse^hausse:1}},
PP(2:3){PREP{des^de:2},NP(3:3){NOUN{prix^prix:3}}},PP(4:6){PREP{de^
de:4},NP(5:6){DET{l'^le:5},NOUN{aluminium^aluminium:6}}},FV(7:7){VERB
{réduira^réduire:7}}},NP(8:9){DET{les^le:8},NOUN{marges^marge:9}},
PP(10:11){PREP{des^de:10},NP(11:11){NOUN{fabricants^fabricant:11}}},
PP(12:13){PREP{de^de:12},NP(13:13){NOUN{boîtes^boîte:13}}},SENT{.^.:14}}
```

3. Affichage simplifié avec des lemmes et non pas des formes de surface :

```
0>GROUPE(0:14){SC(0:7){NP(0:1){DET{le},NOUN{hausse}},PP(2:3){PREP{de},
NP(3:3){NOUN{prix}}},PP(4:6){PREP{de},NP(5:6){DET{le},NOUN{aluminium}}},
FV(7:7){VERB{réduire}}},NP(8:9){DET{le},NOUN{marge}},PP(10:11){PREP{de},
NP(11:11){NOUN{fabricant}}},PP(12:13){PREP{de},NP(13:13){NOUN{boîte}}},
SENT{.}}
```

4. Affichage réduit, uniquément avec des informations syntaxiques :

```
0>GROUPE{SC{NP{La hausse} PP{des NP{prix}} PP{de NP{l' aluminium}}
FV{réduira}} NP{les marges} PP{des NP{fabricants}} PP{de NP{boîtes}} .}
```

B.3 Analyse syntaxique : extraction de dépendances

Liste de dépendances potentiellement fournies par XIP-F.

- `SUBJ` (sujet) ;
- `DEEPSUBJ` (sujet profond) ;
- `VARG` (argument verbal) ;
- `VMOD` (modifieur verbal) ;
- `NARG` (argument nominal) ;
- `NMOD` (modifieur nominal) ;
- `AUXIL` (auxiliaire) ;
- `DETD` (déterminant) ;
- `QUANTD` (quantifieur) ;
- `CONNECT` (connecteur) ;
- `NEGAT` (négation).

Traits associés aux dépendances :

- `NOUN` (nom) ;
- `PRON` (pronom) ;
- `ADJ` (adjectif) ;
- `ADV` (adverbe) ;
- `DIR` (directe) ;
- `INDIR` (indirecte) ;
- `REL` (relatif) ;
- `SPRED` (prédicatif) ;
- `PARTIT` (partitif) ;
- `POSIT1` (première position) ;
- `POSIT2` (deuxième position) ;
- `POSIT3` (troisième position) ;
- `RIGHT` (droit) ;

- LEFT (gauche) ;
- COREF (co-référence).

Sorties pour la phrase en exemple :

La hausse des prix de l'aluminium réduira les marges des fabricants de boîtes .

1. Affichage simplifié tout en conservant des informations morphologiques :

```
SUBJ_NOUN(<réduira^réduire:7>,<hausse^hausse:1>)
VARG_DIR_NOUN(<réduira^réduire:7>,<marges^marge:9>)
NARG_INDIR_NOUN(<hausse^hausse:1>,<des^de:2>,<prix^prix:3>)
NARG_INDIR_NOUN(<marges^marge:9>,<des^de:10>,<fabricants^fabricant:11>)
NARG_INDIR_NOUN(<prix^prix:3>,<de^de:4>,<aluminium^aluminium:6>)
NARG_INDIR_NOUN(<fabricants^fabricant:11>,<de^de:12>,<boîtes^boîte:13>)
```

2. Affichage réduit :

```
SUBJ_NOUN(réduire,hausse)
VARG_DIR_NOUN(réduire,marge)
NARG_INDIR_NOUN(hausse,de,prix)
NARG_INDIR_NOUN(marge,de,fabricant)
NARG_INDIR_NOUN(prix,de,aluminium)
NARG_INDIR_NOUN(fabricant,de,boîte)
```

Annexe C

Scripts de pré-traitements

C.1 Script de balisage de listes

Le fragment suivant correspond à l'identification d'une amorce. Les commentaires apparaissent, en `Perl`, après le symbole `#`.

```
1 $file = shift(@ARGV);                    # mise fichier dans variable

2 open(FILE1,"$file") || die ("cannot open input file $file"); # ouverture fichier
3 open(FILE2, ">$file.bal");                                   # fichier de sortie

4 @lin =<FILE1>;     # liste de lignes du fichier
5 $cnt = 0;          # compteur de lignes du fichier
6 $listes = 0;       # compteur de listes
7 $long = @lin;      # longueur de la liste/fichier (nbre de lignes)
8 $linprec = "";     # ligne précédente

9 while ($cnt<$long) {                      # lignes du fichier à traiter

10 if (($lin[$cnt] =~ /\:([^a-zA-Z])\Z/) && ($lin[$cnt] =~ /[0-9a-zA-z]+/)) {
                                                  # si une ligne finit par :
11   if ($lin[$cnt] =~ /(\:)\s*$/) {              # mise des : dans variable
12    $sepam = $1;
13    $sepam =~ s/\:/\<sp\>\:\<\/sp\>/;            # attribution balise <sp>:</sp>
14   }

15   chop ($lin[$cnt]);
16   if ($lin[$cnt] =~ /\:\s*$/) {
17     $lin[$cnt] =~ s/\://;
18   }

19   $lin[$cnt] =~ s/\./ \<sp\>\.\<\/sp\> /g;   # balise <sp>.</sp>
20   $lin[$cnt] =~ s/\:/ \<sp\>\:\<\/sp\> /g;   # balise <sp>:</sp>
```

```
21   $lin[$cnt] =~ s/\!/ \<sp\>\!\<\/sp\> /g;   # balise <sp>!</sp>
22   $lin[$cnt] =~ s/\?/ \<sp\>\?\<\/sp\> /g;   # balise <sp>?</sp>

23   if ($amorce ne "") {                       # fausses amorces
24   $amorce  =~ s/\<sp\>\:\<\/sp\>/:/ ;        # effacer faux <sp>:</sp>
25   print FILE2 "$amorce\n\n";
26   }                                          # lignes finissant par :

27    @wordlist = split(/ /, $lin[$cnt]);       # comptage des mots de l'amorce
28    for ($i=0; $i<=$#wordlist; $i++) {
29      if ($wordlist[$i] =~ /[0-9a-zA-Z]+/) {
30        $motsamorce++;
31      }
32    }

33    if ($motsamorce > 1) {             # pour être amorce > 1 mot
34 $amorce = $lin[$cnt] . " ";
35 $amorce = $amorce . $sepam;           # concatenation des mots de l'amorce
36 $listes++;                            # comptage de la liste
37    }

38    if ($motsamorce == 1) {
39 $list = 0;
40 print FILE2 "$lin[$cnt] :\n";         # fausses amorces d'1 seul mot
41    }

42    $motsamorce = 0;
43    $linprec = $lin[$cnt];
44    $cnt++;
45  }
46 }
```

C.2 Script de balisage des titres

```
1 $file = shift(@ARGV);                    # mise fichier dans variable

2 open(FILE1,"$file") || die ("cannot open input file $file"); # ouverture
3 open(FILE2, ">$file.tit");                                   # fichier de sortie

4 @lin =<FILE1>;     # liste de lignes du fichier
5 $long = @lin;      # longueur de la liste/fichier (nbre de lignes)
6 $cnt = 0;          # compteur de lignes du fichier
7 $tit = "";         # ligne de titre
```

```
8 $cptetit = 0;      # compteur de titres

9 while ($cnt<$long) {                     # lignes du fichier à traiter

    # lignes suposées titres
    # lignes ne finissant pas par . ou < > mais finissant par caractère
10  if (($lin[$cnt] !~ /\.([^a-zA-Z])\Z/) &&
    ($lin[$cnt] !~ /\</) && ($lin[$cnt] =~ /[0-9a-zA-z]+/)) {
11     $tit = $lin[$cnt];

       # lignes finissant par :  .(1)  .>>  ."  .]  .)
12     if ( ($tit !~ /\:([^a-zA-Z])\Z/) && ($tit !~ /\.(\s)\([0-9]+\)\Z/) &&
   ($tit !~ /\.(\s)+[ż|\"|\]|\)]\Z/) && ($tit !~ /\!(\s)+\Z/) &&
   ($tit !~ /\?(\s)+\Z/)) {

13     $tit =~ s/\./ \<sp\>\.\<\/sp\> /g;   # balise <sp>.</sp>
14     $tit =~ s/\:/ \<sp\>\:\<\/sp\> /g;   # balise <sp>:</sp>
15     $tit =~ s/\:/ \<sp\>\;\<\/sp\> /g;   # balise <sp>;</sp>
16     $tit =~ s/\!/ \<sp\>\!\<\/sp\> /g;   # balise <sp>!</sp>
17     $tit =~ s/\?/ \<sp\>\?\<\/sp\> /g;   # balise <sp>?</sp>
18     print FILE2 "<tit> $tit </tit>\n";   # balises <tit> et </tit>
19     $cptetit++;                          # comptage de titres
20   }

21     else {
22 print FILE2 "$tit";                         # impression sans balises
23     }                                       # ce n'est pas un vrai titre
24    $cnt++;
25  }

26 elsif ($lin[$cnt] !~ /[0-9a-zA-Z]+/) {      # lignes blanches (espaces verticaux)
27    print  FILE2 "$lin[$cnt]";
28    $tit = "";
29    $cnt++;
30  }

31 else {                                      # autres lignes du fichier
32    print  FILE2 "$lin[$cnt]";
33    $tit = "";
34    $cnt++;
35  }
36 }
```

Annexe D

Grammaires de notre modèle

Les sections suivantes (à l'exception de D.4 et D.5 qui donnent des spécifications sur les dépendances) présentent différents types de règles constituant nos grammaires. La typologie, la syntaxe et la sémantique des règles ont été définies dans le chapitre 4 (section 4.5).

Il est à signaler que dans le formalisme XIP les commentaires des règles se présentent entre les symboles / et \.

D.1 Règles de la grammaire noyau

Échantillon des règles de *chunking* pour le marquage de segments additionnels et phrases noyau dans notre modèle.

```
/Segment post. au NP sujet (PNP)\

8>PNP -> |np| np[first:+], (adv+), (pp+), (np+), (iv), (gv),
         (verb[part:+,pas:+]), (ap+), (negpas), (conj[form:fcomme]),
         (sbc)  |fv[verb:+,fin:+]|.

8>PNP -> |np| pp[first:+], (adv+), (pp+), (np+), (iv), (gv),
         (verb[part:+,pas:+]), (ap+), (negpas), (conj[form:fcomme]),
         (sbc) |fv[verb:+,fin:+]|.

8>PNP -> |np| gv[first:+], (adv+), (negpas), (np), (pp+), (np),
         (verb[part:+,pas:+]), (ap+), (sbc), (conj[form:fcomme]),
         (iv), (gv) |fv[verb:+,fin:+]|.

8>PNP -> |np| verb[first:+,part:+,pas:+], (pp+), (adv+), (negpas),
         (ap+), (np), (sbc), (conj[form:fcomme]),(iv), (gv) |fv[verb:+,fin:+]|.


/Segment ant. au NP sujet (ANP)\
```

```
9>ANP[start:+] -> np[time:+,start:+], (adv+), (negpas+), (pp+), (np),
                  punct[last:+,form:fcm].

9>ANP[start:+] -> pp[start:+], (adv+), (negpas+), (pp+), (ap+), (gv+), (np),
                  (verb[part:+,pas:+]), (sbc), punct[last:+,form:fcm].

9>ANP[start:+] -> adv[start:+], punct[form:fcm], (pp), (ap+), (adv+),
                  punct[last:+,form:fcm].

9>ANP[start:+] -> gv[start:+], (np), (adv+), (negpas+), (pp+), (ap+),
                  (verb[part:+,pas:+]), punct[last:+,form:fcm].

9>ANP[start:+] -> iv[start:+], (np), (adv+), (negpas+), (pp+), (ap+),
                  (verb[part:+,pas:+]), punct[last:+,form:fcm].


/Segment post. au FV principal (PFV)\

11>PFV -> |fv| np, (adv*), (negpas*), (np*), (ap*), (pp*), (gv*), (iv*),
          (verb[part:+,pas:+]*), (punct[form:fcm]), (punct[form:fcm]), (sbc),
          (bg) |~fv|.

11>PFV -> |fv| pp, (adv*), (negpas*), (np*), (ap*), (pp*), (gv*), (iv*),
          (verb[part:+,pas:+]*), (punct[form:fcm]), (punct[form:fcm]), (sbc),
          (bg) |~fv|.

11>PFV -> |fv| ap, (adv*), (negpas*), (np*), (ap*), (pp*), (gv*), (iv*),
          (verb[part:+,pas:+]*), (punct[form:fcm]), (punct[form:fcm]), (sbc),
          (bg) |~fv|.


/Phrases de niveau 1\

/Structure de base SVO\

13>S  =  (ANP), np, (PNP),
        fv[verb:+], (PFV),
        sent[form:~fexclm,form:~fintrg].

/Coordination de NPs sujet\

13>S  = |?*[form:~fcm]| np, (ap), coord, np, (PNP),
        fv[verb:+], (PFV),
        sent[form:~fexclm,form:~fintrg].
```

```
/Coordination de verbes finis\

13>S  = (ANP), np, (PNP),
        fv[verb:+], PFV, coord, fv, (PFV),
        sent[form:~fexclm,form:~fintrg].
```

D.2 Règles des grammaires de découpage spécialisées

Échantillon des règles de *chunking* pour le marquage de la ponctuation (guillemets et parenthèses), des listes et des énumérations.

```
/Guillemets\

21>GM = punct[form:foquotes], ?*[form:~fcquotes], punct[form:fcquotes].

22>GM = punct[form:foquotes], ?*[form:~fcquotes], sent[form:~fexclm,form:~fintrg].

23>GM = punct[form:fquotes], ?*[form:~fquotes], punct[form:fquotes].

24>GM = ?[form:~fquotes,start:+], ?*[form:~fquotes], punct[form:fquotes],
punct[form:fcm] |fv|.


/Parenthèses\

27>PR = punct[form:fopar], ?*, punct[form:fcpar].

28>PR = punct[form:fopar], ?*, sent[form:~fexclm,form:~fintrg].


/Correction erreurs de chunking\

30>NP  = NP[first:+,det:+,noun:~], GM.

31>PP  = prep[prep:+,det:+], GM{np[noun:+]}.

31>PP = prep[prep:+,form:fen,sfen:+], GM{np[noun:+]}.


/Amorces des listes\

40>AMR = | sgml[start:+,punct:+,sgml:+,startbis:+] |
```

```
         ?*[sgml:~,form:~fpfin], punct[sgml:+,form:f2pts].

41>AMR = | punct[sgml:+,form:fpfin] |
         ?*[sgml:~,form:~fpfin], punct[sgml:+,form:f2pts].

41>AMR = | punct[sgml:+,form:fintrg] |
         ?*[sgml:~,form:~fpfin], punct[sgml:+,form:f2pts].


/Items des listes\

41>IT = | AMR | ?+[item:~,sgml:~], punct[sgml:+].

42>IT =  ?+[item:~,sgml:~,amorce:~,form:fdebtit], punct[sgml:+] | ~AMR | .

43>IT = | IT | ?+[item:~,sgml:~], punct[sgml:+].


/Enumérations à 3 items\

8>ENM ->  |?[fin:~], conj[form:fcomme]|
        np[first:+,time:~,pron:~], punct+[form:fcm,first:~,last:~],
        np+[time:~,pron:~], adv*, iv*, gv*, ap*, pp*, verb*[partpas:+],
        punct*[form:fhyph], (sbc), coord+[first:~,last:~].

8>ENM -> |punct[form:f2pts]|
        np[first:+,time:~,pron:~], punct+[form:fcm,first:~,last:~],
        np+[time:~,pron:~], adv*, iv*, gv*, ap*, pp*, verb*[partpas:+],
        punct*[form:fhyph], (sbc), coord+[first:~,last:~].

8>ENM -> |?[coord:~]|
        np[first:+,time:~,pron:~], punct[form:fcm,first:~,last:~],
        punct+[form:fcm,first:~,last:~], np+[time:~,pron:~], ap*, pp*,
        verb*[partpas:+], punct*[form:fhyph], adv[etc:+].

8>ENM -> |?[coord:~]|
        np[first:+,time:~,pron:~], punct[form:fcm,first:~,last:~],
        punct+[form:fcm,first:~,last:~], np+[time:~,pron:~], ap*, pp*,
        verb*[partpas:+], punct*[form:fhyph], punct[form:f3p].

8>ENM -> |punct[form:fopar]|
        np+[time:~,pron:~], punct+[form:fcm,first:~,last:~], ap*, pp*,
        punct*[form:fhyph], punct[form:f3p] |punct[form:fcpar]|.

8>ENM -> |punct[form:fhyph]|
```

```
        np+[time:~,pron:~], punct+[form:fcm,first:~,last:~], ap*, pp*,
        punct[form:f3p] |punct[form:fhyph]|.

8>ENM -> |?[coord:~]|
        np[first:+,time:~,pron:~], punct[form:fcm,first:~,last:~],
        punct+[form:fcm,first:~,last:~], np+[time:~,pron:~], ap*, pp*,
        verb*[partpas:+], punct*[form:fhyph]  | sent |.
```

D.3 Règles des grammaires de dépendances spécialisées

Échantillon des règles d'extraction de dépendances pour quelques phénomènes des phrases de niveau 2.

```
/OBJ entre V et " "\

OBJGM(#1,#2) =  FV{?*,#1[last:+,form:~fetre,argspred:~]}, (PP), GM#2.

OBJGM(#1,#2) =  SBC{?*, FV{?*,#1[last:+,form:~fetre,argspred:~]}}, (PP), GM#2.

OBJGM(#1,#2,#3) =  FV{?*,#1[last:+,form:~fetre,argspred:~]}, bg#2,
                GM#3{?,?*[verb:+],?[verb:~,last]}.


OBJGM(#1,#2) =  GV{?*,#1[last:+,form:~fetre,argspred:~]}, GM#2.

OBJGM(#1,#2) =  GV{?*,#1[last:+,form:~fetre,argspred:~]}, NP{?*[det:+],GM#2}.

OBJGM(#1,#2) =  IV{?*,#1[last:+,form:~fetre,argspred:~]}, GM#2.

OBJGM(#1,#2) =  IV{?*,#1[last:+,form:~fetre,argspred:~]}, NP{?*[det:+],GM#2}.


OBJGM(#1,#2) =  #1[part:+,pas:+,partpas:+], GM#2.

OBJGM(#1,#2) =  GM#2{[start:+]}, ?*[punct:+,form:fcm],
                FV{?*,#1[last:+,form:~fetre,argspred:~]}, NP,
                sent[form:~fexclm,form:~fintrg].

OBJGM(#1,#2) = FV{?*,#1[last:+,form:~fetre,argspred:~]},
               punct[form=f2pts,strongbreak=+], GM#2.

OBJGM(#1,#2) = SBC{?*, FV{?*,#1[last:+,form:~fetre,argspred:~]}}, PR, GM#2.
```

```
/NPR entre N et () contigus\

NPR(#1,#2) = NP{?*,#1[last:+]}, PR#2{?,?*[punct:~,last]}.

NPR(#1,#2) = NP{?*,#1[last:+]}, PR#2.


/NPR entre N et () non contigus\

NPR(#1,#2) = NP{?*[det:+],#1[last:+]}, AP, PP, PR#2.

NPR(#1,#2) = NP{?*[det:+],#1[last:+]}, PP, PP, PR#2.

NPR(#1,#2) = NP{?*,#1[last:+]}, NP{?[first:+,last:+,num:+]}, PR#2.

NPR(#1,#2) = NP{#1[pron:~]}, FV, ?*[noun:~, pron:~, prep:~], PR#2.

NPR(#1,#2) = PP{?*[prep:+], NP{?*,#1[last:+]}}, (AP),
             PR#2{?,[punct:~,form:~f3p],?}.

NPR(#1,#2) = NP{?*,#1[last:+]}, PP, PR#2{?,[punct:~,form:~f3p],?}.

NPR(#1,#2) = NP{?*,#1[last:+]}, GM, PR#2{?,[punct:~,form:~f3p],?}.

NPR(#1,#2) = GM#1, PR#2.


/NGM entre N et N ou Adj entre " "\

NGM(#1,#2) = NP{?*[det:+],#1[last:+]}, prep[form:fde],
                GM#2{?,?*[verb:~],?[verb:~,last]}.

NGM(#1,#2) = NP{?*[det:+],#1[last:+]}, ?+[part:+,pas:+,partpas:+],
                GM#2{?,?*[verb:~],?[verb:~,last]}.

NGM(#1,#2) = PP{?*[prep:+],NP{?*[det:+],#1[last:+,pron:~]}}, (AP),
                GM#2{?,?*[verb:~],?[verb:~,last]}.

NGM(#1,#2) = PP{?*[prep:+],NP{?*,#1[last:+,proper:+]}}, (AP),
                GM#2{?,?*[verb:~],?[verb:~,last]}.
```

```
NGM(#1,#2) = NP{?*,#1[last:+]},
                PP{?*[prep:+,form:fde],GM#2{?,?*[verb:~],?[verb:~,last]}}.

NGM(#1,#2) = GM#1[start:+], NP{?*,#2[last:+]}.

NGM(#1,#2) = PR#1, GM#2.

NGM(#1,#2) = NP{?*,#1[last:+]}, (PP), (AP),
             FV{?*,[last:+,form:fetre,argspred:~]}, (NP), punct[form:f2pts], GM#2.

NGM(#1,#2) = NP{?*,#1[last:+]}, (AP), GM#2.

NGM(#1,#2) = NP{?*,#1[last:+]}, (AP), GM, coord, GM#2.

NGM(#1,#2) = PP{?*[prep:+],NP{?*,#1[last:+]}}, PP{?*[prep:+],GM#2}.



/Items des listes\

ITMLIST[dir=+](#1) = SGML, ?*, IT{?*, SP, NP{?*,#1[last:+,pron:~,dem:~]}}.

ITMLIST(#1) = SGML, ?*, IT{?*, SP, PP{?*[prep:+,form:fde],NP{?*[det:+],#1[last:+]}}}.

ITMLIST(#1) = SGML, ?*, IT{?*, SP, PP{?*[prep:+,form:fa],NP{?*[det:+],#1[last:+]}}}.

ITMLIST(#1) = SGML, ?*, IT{?*, SP, AP{?*,#1[last:+]}}.


ITMLIST(#1) =  AMR{?*, punct[sgml:+,form:f2pts]},
               IT{NP{?*,#1[last:+,pron:~,dem:~]}}.

ITMLIST(#1) =  IT{?*, punct[sgml:+,form:fptvirg]},
               IT{NP{?*,#1[last:+,pron:~,dem:~]}}.


if ( ~ITMLIST[dir](?,?) )
ITMLIST(#1) =  AMR{?*, punct[sgml:+,form:f2pts]}, ?*,
                IT{?*, punct[sgml:+,form:fpfin]},
                IT{NP{?*,#1[last:+,pron:~,dem:~]}}.


ITMLIST(#1) =  AMR{?*,punct[sgml:+,form:f2pts]}, IT{AP{?*,#1[last:+]}}.

ITMLIST(#1) =  IT{?*,punct[sgml:+,form:fptvirg]}, IT{AP{?*,#1[last:+]}}.
```

```
ITMLIST(#1) =  IT{?*,punct[sgml:+,form:fpfin]}, IT{AP{?*,#1[last:+]}}.

ITMLIST(#1) =  IT{?*,punct[sgml:+,form:fvirg]}, IT{AP{?*,#1[last:+]}}.



/Clé dans titres\

CLE(#1) = punct[form:fdebtit,sgml:+,start:+,first:+], SP*,
          NP{?*, #1[last:+,pron:~,num:~]}.
```

D.4 Spécification des dépendances de base

- SUBJ : sujet d'un `FV`, `IV` ou `GV` (actifs, passifs et réflexifs).
  `SUBJ_REFLEXIVE(écroule,monde)`
  "Ton monde s' écroule autour de toi."

  `SUBJ(avaient,parents)`
  `SUBJ_PASSIVE(touché,Elgersma)`
  "Par conséquent, Walter Elgersma était directement touché par la loi en question et ses parents avaient un intérêt véritable dans le résultat de cette loi."

- COMP : complément objet d'un `FV`, `IV` ou `GV`
  `COMP_OBJ(représente,marché)`
  "Elle représente un marché en croissance."

  `COMP_OBJ(ausculter,état)`
  "Il est venu en Europe notamment pour ausculter l'état de santé de ce couple vedette des marchés des changes."

- VN : attribut nominal d'un verbe prédicatif.
  `VN(est,exérèse)`
  "Le traitement de principe de la mélanose de Dubreuilh est l'exérèse chirurgicale."

- VADJ : attribut adjectif d'un verbe prédicatif.
  `VADJ(est,forte)`
  "La concurrence est forte dans le domaine du commerce alimentaire."

- VPP : complément prépositionnel d'un `FV`.
  `VPP(adaptée,en,fonction)`
  "La surveillance sera adaptée en fonction traitement institué."

- VAG : agent d'un `FV` passif.
  `VAG(critiquée,par,Guigou)`
  "Elle avait été vivement critiquée par Elisabeth Guigou pour sa gestion de l'enquête sur la Scientologie."

- VADV : complément adverbial d'un `FV`.
  `VADV(repose,essentiellement)`
  "Ce volet repose essentiellement sur deux dispositifs."

- NN : complément nominal d'un `NP`.
  `NN(version,papier)`
  "Ce livre est sorti simultanément en version papier et en version électronique."

- NPP : complément prépositionnel d'un `NP`.
  `NPP(manque,de,transparence)`
  "Il y a d'abord un manque de transparence."

- NADJ : complément adjectif suivant un `NP`.
  `ADJ(témoin,lumineux)`
  "En même temps que le phare est allumé, le témoin lumineux correspondant sur le tableau de bord s'allume."

- ADJN : complément adjectif précédant un `NP`.
  `ADJN(plein,jour)`
  "Un motocycliste prudent tient compte du fait qu'il doit signaler sa présence aux autres conducteurs, même en plein jour.

- COORDFV : coordination non ambiguë entre deux verbes.
  `COORDFV(est,vise)`
  "La liberté d'association est de nature publique et collective et ne vise pas les relations familiales."

- COORDNP : coordination non ambiguë entre deux noms.
  `COORDNP(Protection,assistance)`
  "Protection et assistance sont fournies à la famille au Canada, grâce à différentes mesures législatives et administratives."

D.5 Spécification des dépendances pour les listes

- `SUBJLIST`
  Les items sont les sujets inversés du verbe principal de l'amorce.

  " Un décret en Conseil d'Etat détermine les conditions dans lesquelles sont fixées :

La liste limitative des espèces animales non domestiques (...) ;
La durée des interdictions permanentes ou temporaires (...) ;
La réglementation de la recherche (...).”

```
SUBJLIST(fixées,liste)
SUBJLIST(fixées,durée)
SUBJLIST(fixées,réglementation)

50>MAX{<list> AMR{NP{Un décret} PP{en NP{Conseil d'Etat}} FV{détermine}
NP{les conditions} SBC{BG{PP{dans lesquelles}} FV{sont fixées}} <sp>:</sp>}

IT{NP{La liste} AP{limitative} PP{des NP{espèces}} AP{animales} AP{non
domestiques} (...) IT{NP{La durée} PP{des NP{interdictions}} AP{permanentes
ou temporaires} (...) IT{NP{La réglementation} PP{de NP{la recherche}}
<sp>.</sp>} </list>}
```

– VARGLIST
Les items sont des arguments d'un verbe dans l'amorce requérant un syntagme prépositionnel. La préposition peut apparaître comme dernier élément dans l'amorce ou peut se répéter en début de chaque item.

“ L'évaluation des conséquences économiques de la non-qualité repose principalement sur :
Le coût de l'obtention de la qualité (...)
Les pertes de recettes dues aux écarts qualité (...)”

```
VARGLIST(repose,sur,coût)
VARGLIST(repose,sur,pertes)

231>MAX{<list> AMR{NP{L' évaluation} PP{des NP{conséquences}} AP{économiques}
PP{de NP{la non-qualité}} FV{repose} principalement sur <sp>:</sp>}

IT{SP{1 <sp>.</sp>} NP{Le coût} PP{de NP{l' obtention}} PP{de NP{la qualité}}
(...) <sp>.</sp>} IT{SP{2 <sp>.</sp>} NP{Les pertes} PP{de NP{recettes}} AP{dues}
PP{aux NP{écarts}} NP{qualité} (...) <sp>.</sp>} </list>}
```

– COMPLIST
Les items sont des objets directs du verbe principal de l'amorce. On déduit les compléments étant des syntagmes nominaux mais pas les propositions complétives (*que-phrases*).

“ Les principales critiques ont concerné :
- les données épidémiologiques,
- les examens complémentaires du bilan initial,
- la place de la dermatoscopie,

- les marges d'exérèse. "

```
COMPLIST(concerné,données)
COMPLIST(concerné,examens)
COMPLIST(concerné,place)
COMPLIST(concerné,marges)

1>MAX{<list> AMR{NP{Les AP{principales} critiques} FV{ont concerné}
<sp>:</sp>}

IT{SP{-} NP{les données} AP{épidémiologiques} <sp>,</sp>} IT{SP{-}
NP{les examens} AP{complémentaires} PP{du NP{bilan}} AP{initial}
<sp>,</sp>} IT{SP{-} NP{la place} PP{de NP{la dermatoscopie}} <sp>,
</sp>} IT{SP{-} NP{les marges} PP{d' NP{exérèse}} <sp>.</sp>} </list>}
```

– VNLIST et VADJLIST
  Les items sont des attributs nominaux ou adjectifs d'un verbe prédicatif.

  " Les deux segments visés sont :
  - commercial (hôtels, restaurants, restauration rapide, stations de vacances) .
  - institutionnel (hôpitaux, établissements carcéraux, maisons d'enseignement, sociétés minières). "

```
VADJLIST(sont,commercial)
VADJLIST(sont,institutionnel)

159>MAX{<list>AMR{NP{Les deux segments} NP{visés} FV{sont} <sp>:</sp>}

IT{AP{commercial} PR{( NP{hôtels} , NP{restaurants} , NP{restauration}
AP{rapide} , NP{stations} PP{de NP{vacances}} )} <sp>.</sp>} IT{AP{
institutionnel} PR{( NP{hôpitaux} , NP{établissements} AP{carcéraux},
NP{maisons} PP{d' NP{enseignement}} , NP{sociétés} AP{minières} )}
<sp>.</sp>} </list>}
```

D.6 Règles pour le rattachement prépositionnel

Échantillon de règles « fiables » (MF1) et « non fiables » (MF2), appartenant à la grammaire de base G2, pour l'extraction des dépendances liées au rattachement prépositionnel.

```
/--- MF1 ---\

/1 PP contigu avec même rection\

if (#1[subcatform]:#2[subcatform])
A[mf1=+](#1,#2,#3) =
    FV{?*,#1[last:+]};IV{?*,#1[last:+]};
    GV{?*,#1[last:+]};SBC{?*,FV{?*,#1[last:+]}};verb#1[partpas:+],
    ?+[form:~fcm,form:~fopar,form:~fcpar,form:~f2pts,verb:~,scbegin:~,coord:~],
    (adv[psneg:+]), PP[fonc:~,time:+,fvpp=+,fadjpp:~,!noun:!,!pron:!]
    {?*[prep:~],#2[prep:+,sfde:~],?*,NP{?*,#3[last:+]}}.
```

Exemple :

A_MF1_1(donner lieu,à,sortie)
Ce sont des charges quelque peu particulières, qui ne donnent pas lieu à une sortie de caisse.

```
/2 PP initial\

A[mf1=+](#1,#2,#3) =
    PP[start:+]{?*, #2[prep:+], ?*,NP{?*,#3[last:+]}}, punct[form:fcm],
    ?+[verb:~,svinfdir:~,scbegin:~,coord:~],
    FV{?*,#1[last:+]};IV{?*,#1[last:+]};
    GV{?*,#1[last:+]};SBC{?*,FV{?*,#1[last:+]}};verb#1[partpas:+].
```

Exemple :

A_MF1_2(pouvoir,dans le cadre de,tolérance)
Dans le cadre de cette tolérance, un maximum de 4 % peuvent présenter des germes extérieures visibles.

```
/3 rection avec préposition 'de' lors de NP PP\

if (~A(?,#2,#3))
A[mf1=+](#1,#2,#3) =
    NP[start:~]{?*,#1[last:+,num:~,time:~,pron:~]};
```

```
    PP[start:~]{?*,NP{?*,#1[last:+,num:~,time:~,pron:~]}};
    BG{?*,pron#1[last:+]},
    ?*[noun:~,verb:~,prep:~,pron:~,coord:~,form:~fopar,form:~fcpar,
    form:~f2pts,conjque:~,punct:~],
    PP[fnpp=+]{?*, #2[prep:+,sfde:+], ?*,NP{?*,#3[last:+]}}.
```

Exemple :

A_MF1_3(niveau,de,stocks)
Le niveau de stocks diffère également.

```
/--- MF2 ---\

/Vérification que #3 n'est pas déjà rattaché avec une dépendance [mf1]\

/8 PP non contigu après un autre PP\

if (~A[mf1:+](?,?,#3) )
A[mf2=+](#1,#2,#3) =
    NP{?*,#1[last:+]};PP{?*,NP{?*,#1[last:+]}},
    ?*[verb:~,prep:~,form:~f2pts,conjque:~,form:~fcm,punct:~,coord:~],
    PP, ?*[noun:~,verb:~,prep:~,form:~f2pts,conjque:~,coord:~,form:~fcm],
    PP{?*, #2[prep:+], ?*,NP{?*,#3[last:+]}}.
```

Exemple :

A_MF2_8(résultat,au cours de,années)
Cela n'est pas sans incidence sur le résultat de la société au cours des premières années de son exploitation.

```
/9 PP après adjectif\

if (~A[mf1:+](?,?,#3) )
A[mf2=+](#1,#2,#3) =
    AP{?*,#1[last:+]},
    ?*[verb:~,prep:~,form:~f2pts,conjque:~,coord:~,form:~fcm,adj:~,noun:~],
    PP{?*, #2[prep:+], ?*,NP{?*,#3[last:+]}}.
```

Exemple :

A_MF2_9(inhérentes,aux,licenciements)
Elles sont censées couvrir les charges inhérentes aux licenciements.

```
/12 S'il n'y a pas de relation MF2 déjà extraite avec le premier argument\
/   Rattachement du PP (sauf préposition 'de') au verbe\

if (~A[mf1:+](?,?,#3) && ~A[mf2:+](#1,?,?) )
A[mf2=+](#1,#2,#3) =
FV{?*,#1[last:+,form:~fetre]};IV{?*,#1[last:+,form:~fetre]};
GV{?*,#1[last:+,form:~fetre]};SBC{?*,FV{?*,#1[last:+,form:~fetre]}},
?*[verb:~,form:~f2pts,conjque:~,coord:~,form:~fcm], PP, (ADV),
PP{?*, #2[prep:+,sfde:~], ?*, NP{?*,#3[last:+]} }.
```

Exemple :

A_MF2_12(permettre,en fonction de,distance)

Cette fonction permet un réglage de compensation d'amplitude en fonction de la distance.

Annexe E

Corpus pour l'apprentissage

E.1 Description des corpus A et B

Pour le travail concernant la lexicalisation des grammaires de dépendance, nous avons collecté un ensemble de corpus sur le Web appartenant à trois domaines différents : économique (lexique financier), scientifique (télécommunications, acoustique), et juridique (texte d'un arrêté ministériel).

L'ensemble de ces corpus contient 11.628 mots dont 4.764 appartiennent au domaine économique (41 %) , 5.451 au domaine scientifique (46,8 %) et 1.417 au domaine juridique (12,2 %).

La longueur moyenne des phrases est de 25,3 mots dans le domaine économique, 29,4 dans le domaine scientifique et de 37,6 dans le domaine juridique. La moyenne globale est de 28,8 mots par phrase.

Nous avons divisé ces corpus en deux ensembles contenant chacun des phrases des trois domaines. Cela nous a permis de constituer :

- un corpus A (89 phrases et 2.493 mots)
- un corpus B (337 phrases et 9.339 mots)

Comme nous l'avons précisé dans le chapitre 7 (*cf* 7.3.3), le premier corpus a été utilisé pour le développement de la grammaire de base (G2), le deuxième pour la valider. Le corpus B a aussi été le point de départ de nos expériences d'apprentissage. Il a été analysé avec la grammaire G2, puis les dépendances concernant le rattachement prépositionnel et produites par une règle « peu fiable » (type `MF2`, c'est-à-dire ayant un taux de précision < 93 %) ont été extraites. Elles ont servi de base pour la création de requêtes pour le Web et ont permis la création de la base de données lexicales.

E.2 Échantillon du corpus B

Les fragments qui suivent appartiennent au corpus B une fois prétraité pour identifier les limites des phrases (incluant listes et titres).

<tit> Les charges décaissées </tit> La nature des charges , et leur pondération , va dépendre de l' activité de la société . On peut tenter d' approcher les situations les plus typiques au travers d' exemples concrets . Les achats de matières premières concernent surtout les industries de transformation . Ils peuvent représenter une part importante des coûts . Cela peut occasionner une sensibilité assez forte des résultats aux variations de prix de ces matières premières , et aux mouvements du dollar , qui sert le plus souvent de référence à leur facturation . Ainsi , un producteur de petit appareillage électrique sera touché par une hausse du prix des matières plastiques , qui représentent une part significative du prix de revient (quelques dizaines de pour cent) . Un producteur de papier , s' il n' est pas complètement intégré verticalement , devra subir les hausses et les baisses des prix de la pâte à papier . Un producteur de boîtes de conserves ou de boîtes - boissons sera pénalisé par une hausse des cours de l' aluminium et par une hausse du dollar . De même , la forte consommation de produits énergétiques peut avoir des effets significatifs sur la rentabilité d' une entreprise . Un producteur de ciment ou d' aluminium , activité très consommatrice d' énergie , subira les conséquences des variations du prix du pétrole ou de l' électricité . Quelle que soit la qualité de la gestion de l' entreprise , elle ne pourra se départir des effets prix de ces matières , indispensables à sa production . Les effets , dans un sens comme dans l' autre , peuvent être cumulatifs . La hausse des prix de l' aluminium réduira les marges des fabricants de boîtes , qui seront obligés à leur tour de monter leurs prix et pénaliseront alors certains industriels de l' agro-alimentaire . Parfois , un maillon de la chaîne n' aura pas la possibilité de répercuter les hausses de prix qu' il subit sur ses prix de vente et sera donc victime d' un " effet ciseaux " , dramatique pour ses marges : l' industriel de l' agro-alimentaire pourra par exemple se voir opposer une fin de non-recevoir par un grand distributeur pour l' augmentation de ses tarifs .

(...)

<tit> V <sp>.</sp> DESCRIPTION DES SYSTEMES SATELLITAIRES </tit> Plusieurs systèmes de communication mondiale à base de constellation de satellites sont actuellement en projet . <tit> V - 1 <sp>.</sp> Orbcomm </tit> Orbcomm met actuellement en place une constellation de satellites sur orbite basse (LEO) qui seront utilisés pour prélever des paquets de données , afin de soit les transmettre directement à une station au sol si le satellite est au-dessus , soit les stocker à bord dans un mode " store and forward " puis les télécharger lorsque le satellite est en vue d' une station terrestre . La constellation Orbcomm devrait comporter 28 satellites et être mise en service dans le dernier trimestre 98 . La constellation devrait avoir une couverture mondiale puisque nous avons trouvé des systèmes de communication pouvant être embarqués sur des bateaux afin d' assurer des transmissions de données et une localisation par GPS . Ces systèmes , fabriqués par la société Torrey Science Corporation ont les caractéristiques communes suivantes : la transmission de données est bidirectionnelle avec un taux de transfert de 2400 bits par seconde (bps) pour la liaison montante et 4800 bps pour la liaison descendante . <list> Il existe trois modes de fonctionnement <sp> :</sp> Un mode Sleep permettant une économie d' énergie et attendant un signal de départ <sp>.</sp> La consommation est de l'ordre de quelques micro ampères <sp>.</sp> Un mode Standby permettant la réception en continu du signal satellitaire <sp>.</sp> La consommation est de l'ordre de quelques centaines de milliampères <sp>.</sp> Un

mode Transmit pour la communication satellitaire <sp>.</sp> La consommation est de l'ordre de 3 ampères <sp>.</sp> La puissance requise est de 5 watts minimum <sp>.</sp> </list> Ces équipements sont programmables sur site ou par le réseau de satellites Orbcomm . Les dimensions varient de 25x15x5 cm3 (auxquels il faut rajouter une antenne) à 50x30x10 cm3 comprenant une antenne intégrée pour les communications avec le satellite . Le prix d' un terminal est de l'ordre de $ 200 .

(...)

<tit> Catégorie " I " </tit> Les échalotes classées dans cette catégorie doivent être de bonne qualité . Elles doivent présenter la forme et la coloration typiques de la variété ou du type commercial . <list> Les bulbes doivent être <sp> :</sp> - fermes et consistants <sp>;</sp> - de coloration normale par rapport au type commercial auquel ils appartiennent <sp>;</sp> - dépourvus de tige creuse et résistante <sp>;</sp> - exempts de renflements provoqués par un développement végétatif anormal <sp>;</sp> - pratiquement dépourvus de touffe radiculaire , sauf pour les échalotes rondes et les échalotes grises <sp>.</sp> </list> Ils peuvent présenter de petites fissures sur la pellicule extérieure du bulbe . <tit> Catégorie " II " </tit> Cette catégorie comporte les échalotes qui ne peuvent être classées dans la catégorie " I " mais correspondent aux caractéristiques minimales ci-dessus définies . <list> Les échalotes peuvent comporter les défauts suivants , à condition de garder leurs caractéristiques essentielles de qualité , de conservation et de présentation <sp> :</sp> - forme et coloration non typiques du type commercial <sp>;</sp> - lésions d' origine mécanique cicatrisées et légères meurtrissures non susceptibles de nuire à la conservation <sp>;</sp> - marques légères résultant d' attaques parasitaires ou de maladies <sp>.</sp> </list> <tit> II <sp>.</sp> Dispositions concernant le calibrage </tit> Article . 5 - Le calibrage est déterminé par le diamètre maximal de la section équatoriale . Le diamètre minimal est fixé à 10 mm pour les échalotes grises et à 15 mm pour les autres types . <list> La différence de diamètre entre l' échalote la plus petite et la plus grosse contenues dans un même colis ne doit pas excéder <sp> :</sp> - 10 mm lorsque l' échalote la plus petite a un diamètre compris entre 10 mm inclus et 15 mm exclus <sp>;</sp> - 15 mm lorsque l' échalote la plus petite a un diamètre compris entre 15 mm inclus et 20 mm exclus <sp>;</sp> - 20 mm lorsque l' échalote la plus petite a un diamètre égal ou supérieur à 20 mm <sp>.</sp> </list> <tit> III <sp>.</sp> Dispositions concernant les tolérances </tit> Article . 6 - Des tolérances de qualité et de calibre sont admises dans chaque colis , ou chaque lot pour les échalotes expédiées en vrac , pour les produits non conformes aux exigences de la catégorie indiquée . <tit> A <sp>.</sp> Tolérances de qualité </tit>

E.3 Échantillon analysé

Le fragment qui suit correspond aux six premières phrases de l'échantillon du corpus B analysé par la grammaire G2. À côté de chaque relation de dépendance il apparaît le trait de fiabilité (type de règle qui a produit la dépendance, avec un taux de précision supérieur ou inférieur à 93 %) ainsi que le numéro de règle.

```
8><tit> Les charges décaissées </tit>

A_MF1_3(<activité:13>,<de:14>,<société:16>)
A_MF1_4(<nature:1>,<de:2>,<charge:3>)
A_MF2_5(<dépendre:10>,<de:11>,<activité:13>)
9>La nature des charges , et leur pondération , va dépendre de
l' activité de la société .

A_MF2_5(<approcher:4>,<au travers de:10>,<exemple:11>)
A_MF2_7(<situation:6>,<au travers de:10>,<exemple:11>)
A_MF2_7(<typique:9>,<au travers de:10>,<exemple:11>)
10>On peut tenter d' approcher les situations les plus typiques
au travers d' exemples concrets .

A_MF1_3(<industrie:8>,<de:9>,<transformation:10>)
A_MF1_4(<achat:1>,<de:2>,<matière:3>)
11>Les achats de matières premières concernent surtout les industries
de transformation .

A_MF1_3(<part:4>,<de:6>,<coût:7>)
12>Ils peuvent représenter une part importante des coûts .

A_MF1_3(<sensibilité:4>,<de:7>,<résultat:8>)
A_MF1_3(<plus:27>,<de:29>,<référence:30>)
A_MF1_3(<variation:10>,<de:11>,<prix:12>)
A_MF1_3(<prix:12>,<de:13>,<matière:15>)
A_MF1_3(<mouvement:20>,<de:21>,<dollar:22>)
A_MF2_7(<résultat:8>,<à:9>,<variation:10>)
A_MF2_7(<référence:30>,<à:31>,<facturation:33>)
A_MF2_8(<sensibilité:4>,<à:9>,<variation:10>)
A_MF2_8(<plus:27>,<à:31>,<facturation:33>)
A_MF2_12(<occasionner:2>,<à:9>,<variation:10>)
A_MF2_12(<servir:25>,<à:31>,<facturation:33>)
13>Cela peut occasionner une sensibilité assez forte des résultats
aux variations de prix de ces matières premières , et aux mouvements
du dollar , qui sert le plus souvent de référence à leur facturation .
```

E.4 Échantillon corrigé

Le fragment qui suit correspond aux six premières phrases de l'échantillon du corpus B, après correction manuelle à partir d'une première analyse (seulement les dépendances jugées correctes apparaîssent).

```
8><tit> Les charges décaissées </tit>

A(<dépendre:10>,<de:11>,<activité:13>)
A(<activité:13>,<de:14>,<société:16>)
A(<nature:1>,<de:2>,<charge:3>)
9>La nature des charges , et leur pondération , va dépendre de
l' activité de la société .

A(<approcher:4>,<au travers de:10>,<exemple:11>)
10>On peut tenter d' approcher les situations les plus typiques
au travers d' exemples concrets .

A(<industrie:8>,<de:9>,<transformation:10>)
A(<achat:1>,<de:2>,<matière:3>)
11>Les achats de matières premières concernent surtout les industries
de transformation .

A(<part:4>,<de:6>,<coût:7>)
12>Ils peuvent représenter une part importante des coûts .

A(<sensibilité:4>,<de:7>,<résultat:8>)
A(<variation:10>,<de:11>,<prix:12>)
A(<prix:12>,<de:13>,<matière:15>)
A(<mouvement:20>,<de:21>,<dollar:22>)
A(<servir:25>,<de:29>,<référence:30>)
A(<résultat:8>,<à:9>,<variation:10>)
A(<référence:30>,<à:31>,<facturation:33>)
13>Cela peut occasionner une sensibilité assez forte des résultats
aux variations de prix de ces matières premières , et aux mouvements
du dollar , qui sert le plus souvent de référence à leur facturation .
```